UMA ABORDAGEM BASEADA EM RELAÇÕES LÉXICAS PARA AUTORIA DE HIPERTEXTOS

José Maria Nazar David

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS À OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Marcos Roberto da Silva Borges, Ph. D.

(Presidente)

Profa. Sueli Bandeira Teixeira Mendes, Ph. D.

Prof. Fernando Silva Pereira Manso, Ph. D.

RIO DE JANEIRO, R.J. - BRASIL

NOVEMBRO DE 1991

DAVID, JOSÉ MARIA NAZAR
Uma Abordagem Baseada em Relações Léxicas para Autoria de Hipertextos [Rio de Janeiro] 1991
ix, 120 p., 29,7 cm. (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 1991)
Tese - Universidade Federal do Rio de Janeiro, COPPE
1. Hipertextos 2. Autoria 3. Relações Léxicas
3. Coesão Textual
I. COPPE/UFRJ II. Título (série).

Ao meu Pai e

à minha Mãe

Agradecimentos:

Ao Professor Marcos Borges pela hiperorientação, paciência, incentivo e pela experiência transmitida ao longo de todo esse trabalho.

À Professora Sueli Mendes pela co-orientação, dedicação, sugestões e pelo conhecimento transmitido.

Ao Professor Fernando Manso pela participação na banca examinadora desta tese.

Ao meu irmão, Professor Sérgio, pelas inúmeras sugestões e pelo incansável trabalho de revisão do texto.

Aos participantes do grupo de estudos em hipertexto pelas idéias durante nossas reuniões.

Aos meus pais, pela confiança e pelo apoio nos momentos não raros de dificuldades e tensão.

À prima Cláudia pela alegria, carinho e estímulo constante em todos os momentos difíceis.

Ao Programa de Engenharia de Sistemas pelo conhecimento transmitido e por proporcionar condições favoráveis ao desenvolvimento desta tese.

Ao CnPq, CAPES e Sociedade Cultural e Beneficiente Guilherme Guinle pelo apoio financeiro.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para obtenção do grau de Mestre em Ciências (M. Sc.).

Uma Abordagem Baseada em Relações Léxicas para Autoria de Hipertextos

José Maria Nazar David

Novembro de 1991

Orientador: Prof. Marcos Roberto da Silva Borges

Programa: Engenharia de Sistemas e Computação

O processo de geração de ligações constitui um dos problemas mais sérios em hipertextos.

Ferramentas são necessárias para auxiliar o autor nesse processo, principalmente para que se possa reduzir o esforço cognitivo envolvido na geração de ligações relevantes.

Uma ferramenta para auxiliar o estabelecimento de ligações é descrita nesta tese . A ferramenta se baseia na utilização de Relações Léxicas e técnicas de Recuperação da Informação para obter possíveis relacionamentos significativos entre nós.

Abstract of Thesis presented to COPPE/UFRJ as partial fulfillment of the requirements for the degree of Master of Science (M. Sc.)

A Lexical Relation Based Approach to Hypertext Authoring

José Maria Nazar David

November, 1991

Thesis Supervisor: Prof. Marcos Roberto da Silva Borges

Department: Systems Engeneering and Computer Science

The linking process is one of the most serious problems in hypertext systems.

We need tools to assist authoring process, mainly to reduce the cognitive overload when relevant links are established.

A tool to assist linkage is described in this thesis. It's a Lexical Relation and Information Retrieval Techniques based tool to get probable significant relationships between nodes.

ÍNDICE

Capítulo I

Introdução

I.1 - Motivação

O principal objetivo dos sistemas de hipertexto é fornecer mecanismos adequados que possibilitem armazenar e recuperar grandes volumes de informações. Ou seja, mecanismos que possibilitem aos usuários não só constatarem a presença de uma determinada informação como também recuperá-la de forma eficiente. Nesses sistemas tais informações fazem parte de uma rede de nós - correspondendo basicamente a elementos textuais - e ligações que buscam relacioná-los de forma coerente dentro de um contexto.

Inúmeros problemas ainda são encontrados em ambientes de hipertexto, e dentre eles podemos considerar a construção e manutenção de nós e ligações como um dos mais sérios. Ao particionarmos uma idéia para a criação inicial dos nós de uma rede de hipertexto muitas vezes, na finalização, não temos uma visão global do conteúdo da rede suficiente para a criação das ligações. Conseqüentemente, algumas destas ligações, as mais importantes, são abandonadas em função de outras, provavelmente aquelas menos importantes.

Algumas propostas existem na literatura no sentido de desenvolver mecanismos mais inteligentes para o estabelecimento/manutenção de ligações, entretanto poucos trabalhos têm direcionado esforços no sentido de implementar e testar tais

mecanismos. Dentre esses trabalhos, cabe citar o SmarText, desenvolvido pela Lotus Development Corp. (Fersko-Weiss, 1991).

Paralelamente aos estudos realizados na área de hipertexto, muitos pesquisadores têm buscado o estabelecimento de formalismos visando principalmente alcançar níveis adequados de eficiência nos sistemas de recuperação da informação. Apesar dos esforços, muitos trabalhos ainda estão longe de fornecerem métodos e ferramentas mais poderosas para uma completa compreensão do texto. é bem verdade que tais ferramentas são necessárias para que possamos atingir altos níveis de granularidade nos sistemas.

Como resultado dessas pesquisas, vários autores, dentre eles Salton e Fagan (Cornell University), Sparck Jones (Cambridge), Smadja (Univ. Columbia), e Maarek (IBM Research Center), têm sugerido que as abordagens puramente lingüísticas que buscam uma análise, compreensão e expressão do significado podem ser abandonadas em função de outras mais "mecanicistas" para que possamos conseguir níveis aceitáveis de performance. Assim sendo, métodos léxicos-estatísticos de processamento de linguagem natural têm sido utilizados com freqüência, e os testes também têm indicado que índices razoáveis de efetividade podem ser alcançados.

I.2 - Apresentação do Trabalho

Baseado em alguns estudos realizados tanto na área de Hipertexto quanto na área de Recuperação da Informação o trabalho visa abordar alguns conceitos utilizados nessas áreas, e visa também a aplicação do conceito de **relações léxicas** (sintagmáticas) no auxílio ao processo de autoria em hipertexto.

Termos compostos (relações) que aparecem no texto com freqüência são extraídos como unidades básicas de indexação. A eles são fornecidos pesos adequados, de acordo com a freqüência no documento e também na coleção de documentos. Além de trazerem consigo alguma semântica esses termos são capazes de estabelecer uma "similaridade" entre os textos e com isso permitir não só a construção de ligações entre estes textos dentro de um contexto específico, como também são capazes de fornecer meios para a produção de escritas e leituras adequadas.

I.3 - Estrutura do Trabalho

Este trabalho está estruturado da seguinte forma:

1- Capítulo I : Introdução.

2- Capítulo II : Sistemas de Hipertexto.

3- Capítulo III: Autoria em Hipertexto.

4- Capítulo IV : Recuperação da Informação.

5- Capítulo V : Relações Léxicas.

6- Capítulo VI : Sistema de Autoria e Navegação.

7- Capítulo VII: Conclusões e Sugestões para Trabalhos Futuros.

O capítulo II apresenta a definição de hipertexto e também os principais conceitos que envolvem esses sistemas.

O capítulo III aborda aspectos importantes do processo de autoria em hipertextos que exercem uma forte influência tanto na leitura dos textos quanto na navegação na rede. São apresentados também alguns mecanismos utilizados por importantes sistemas de hipertextos visando auxiliar a recuperação e busca da informação numa rede.

O capítulo IV apresenta a definição de alguns termos encontrados nos sistemas de Recuperação da Informação e que são mencionados, com freqüência, nos capítulos posteriores. Adicionalmente, o capítulo discute aspectos sobre a importância de um termo como unidade de indexação de um texto.

O capítulo V apresenta a definição de relações léxicas e procura justificar a sua utilização como unidades básicas de indexação. Adicionalmente, este capítulo aborda um mecanismo para extrair tais unidades a atribuir um peso a elas.

O capítulo VI descreve o protótipo implementado e os resultados obtidos durante a fase de testes. São apresentadas também algumas conclusões preliminares sobre o processo de autoria em hipertextos.

No último capítulo são apresentadas as conclusões do trabalho realizado. Finalmente são feitas algumas sugestões sobre os trabalhos que ainda devem ser feitos no sentido de aprimorar as técnicas utilizadas pelo protótipo e melhorar o seu desempenho.

Capítulo II

Sistemas de Hipertexto

"The human mind (...) operates by association. Man cannot hoppe fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it. One cannot hope to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage."

(V. Bush, "As We May Think", 1945)

Neste capítulo são apresentados a definição de hipertexto, alguns conceitos básicos relacionados aos sistemas de hipertexto e como se classificam esses sistemas.

II.1 - Hipertexto: Definição

Num texto tradicional, onde o leitor naturalmente é conduzido a realizar uma leitura estritamente seqüencial, páginas são lidas seqüencialmente sem oferecer ao leitor sequer outras formas e opções de leituras que certamente poderiam clarificar alguns conceitos e contribuir para uma melhor reflexão sobre a leitura. Uma forma freqüentemente encontrada na literatura para definir hipertexto é compará-lo a um texto tradicional.

Entretanto muitos livros e artigos atuais já buscam oferecer ao leitor - através de índices remissivos, referências e tabelas de conteúdo, por exemplo - seqüências de leituras que dependem muito do interesse do leitor, da sua formação e especialização. Enfim, já antecipam, de alguma forma, a idéia de um hipertexto: uma leitura não seqüencial (ou "multiseqüencial").

Uma forma bastante clara para ilustrar o modelo conceitual de hipertexto seria através de uma analogia a uma enciclopédia tradicional. Nela, cada texto (nó) contém referências (ligações) a outros textos contidos ou não na própria enciclopédia. Através dessa organização já se buscava, intuitivamente, apresentar ao

leitor uma associação prévia de textos e idéias de uma forma bem coerente.

Em um sistema de hipertexto não há uma ordem única de leitura de um texto. Várias opções são oferecidas aos leitores que determinarão a seqüência a ser seguida no momento da leitura. Ao invés de "virar" uma página, e prosseguir a leitura, o leitor, agora, terá que escolher um botão (uma região da tela sensível ao "mouse"), marcado no texto, que automaticamente o conduzirá a um outro texto ou a outro tipo de informação (sons, imagens gráficas, programas de computador, etc.) que, por sua vez, deverá conter outros novos botões que contribuirão para formar a seqüência de leitura (Fig. II.1) (Nielsen, 1990).

Podemos então, ver hipertexto como um sistema (associativo) de gerenciamento da informação onde os dados são armazenados numa rede de nós. Essas informações, conectadas por ligações, podem ser tipadas, bidirecionais ou com atributos. Ou seja, as ligações unem fragmentos de texto ou imagens gráficas com uma breve explicação a outros textos, formando então novos textos, maiores e mais compreensíveis para o usuário.

Muitos sistemas não limitam a textos o conteúdo em um nó de uma rede de hipertexto. Conseqüentemente eles podem conter textos, gráficos, animação, informação de áudio, ou até mesmo um outro hipertexto como no sistema Trellis (Furuta, 1990), por exemplo.

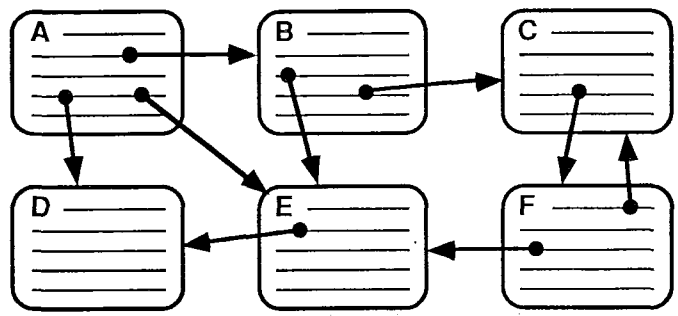


Fig. II.1 – Vista simplificada de uma pequena estrutura um hipertexto tendo seis nós e nove ligações.

Num nível mais básico, FIDERIO (Fiderio, 1988) considera um hipertexto como um Sistema Gerenciador de Banco de Dados (SGBD), que permite associar blocos de informações para vários propósitos, com diferentes estratégias de busca (navegação). Em muitos casos, entretanto, encontramos sistemas de hipertextos que não oferecem mecanismos de busca tão sofisticados como aqueles encontrados nos SGBDs mais recentes. Por exemplo, nos sistemas Hiperties e Intermedia os mecanismos de busca são limitados por "string" (Shneiderman, 1987), (Marchionini, 1988), (Nielsen, 1990).

Num nível mais alto, hipertexto é um ambiente de trabalho cooperativo onde usuários trocam informações, se comunicam e adquirem conhecimento. Compartilhando uma rede de documentos interligados contendo textos ou imagens gráficas, escritores e projetistas podem criar novos documentos e associar idéias

contidas nesses mesmos documentos a outros comentários, notas explicativas ou textos da rede.

Uma boa escrita, especialmente uma escrita técnica, se caracteriza principalmente pela apresentação de várias linhas paralelas de uma história ou um argumento de forma a entrelaçá-las de uma maneira consistente e expressiva (Conklin, 1987). Existem inúmeras ferramentas que auxiliam muito o desenvolvimento de textos e idéias coerentemente. Entretanto elas ainda não fornecem velocidade de acesso à informação, visualização global e compartilhamento do conjunto informações, apresentando o texto como uma seqüência linear.

Textos tradicionais nos conduzem a escrever e ler parágrafos na maioria das vezes como uma sucessão linear, na ordem em que o texto flui, com sinalizações e desvios no desenvolvimento de uma idéia.

Assim sendo, sistemas de hipertexto surgiram principalmente para suprir a necessidade de se organizar fragmentos de textos numa forma não seqüencial (não linear), trazendo consigo todos os mecanismos e facilidades resultantes de tal organização. Dentre tais mecanismos podemos citar aqueles que fornecem ao usuário a possibilidade de estruturar idéias de diferentes formas, de acordo com seus objetivos.

> "(...) A necessidade de um texto não linear não é inerente ao texto, mas relativa a questões ou escolhas que o leitor possa fazer" (Shasha, 1987).

A idéia de um formato não linear de um hipertexto é bastante análoga à maneira como uma pessoa se comporta ao ler um texto ou gráficos. Neles, determinadas palavras ou ilustrações freqüentemente provocam uma associação com outros textos ou ilustrações formando uma rede de idéias onde cada texto contém um "índice associativo".

Produz-se então leitura não só porque o leitor atribui significado, interpreta, mas também porque ele escolhe por onde seu interesse e desejo querem navegar. E o curioso é que, mesmo virando página após página, até que ponto toda leitura não comporta boa margem de escolhas do que se quer ler, até que ponto toda leitura não é aleatória? Seria então toda leitura também uma nova, uma outra, uma escrita segunda?

Apesar de não existirem ferramentas padronizadas, tipicamente os sistemas de hipertexto possuem um editor de texto, um editor gráfico, um banco de dados (onde ainda não existe um consenso em relação ao modelo escolhido) e ferramentas para paginar (que, dentre outras coisas, auxiliam muito no processo de navegação no hiperespaço). Dependendo do tipo de sistema — sistemas específicos a um domínio, por exemplo —, a existência dessas ferramentas poderá diferir em muito.

II.2 - Histórico

O desenvolvimento de novos recursos computacionais e de estações de trabalho mais poderosas, incluindo os vídeos de alta resolução e opções para armazenamento como os CD-ROM, contribuíram de certa forma para que uma antiga idéia de organização da informação - não seqüencial - proposta por VANNEVAR BUSH (Bush, 1945), em 1945, pudesse ser viabilizada. Nessa época Bush propôs a criação de um dispositivo capaz de armazenar livros, fotografias, periódicos, jornais, etc. Através dessa máquina , denominada "memex", o usuário poderia acessar informações com grande velocidade e flexibilidade, adicionar comentários, e criar estruturas associativas que modelam o raciocínio humano.

Desta forma idealizou-se um sistema de informação não-linear possibilitando inclusive que leitores e autores pudessem interagir com o sistema adicionando comentários ou fazendo modificações nos textos em momentos diferentes (sistema de "autoria múltipla"). Na verdade, foram fornecidas as bases fundamentais que posteriormente influenciaram novos pesquisadores na concepção de novos sistemas.

Entretanto nada de prático foi feito com estas idéias durante duas décadas. Em 1963, DOUGLAS ENGELBART no seu artigo " A Conceptual Framework for Augmentation of Man's Intellect" (Engelbart, 1963) escreve que os computadores poderiam determinar

um novo estágio onde indivíduos poderiam se comunicar rapida e facilmente através dele, manipulando símbolos automaticamente.

> "(...) os símbolos com os quais o ser humano representa os conceitos que está manipulando podem ser rearranjados, movidos, armazenados, recuperados e operados de acordo com regras extremamente complexas - tudo em resposta muito rápida a um mínimo de informação fornecida pelo homem (...)"
>
> (Engelbart, 1963)

Em 1968, D. ENGELBART pesquisando o uso dos computadores no desenvolvimento do intelecto humano desenvolve o sistema NLS ("on-line system"). Nesse sistema ele implementa um ambiente de trabalho no qual o usuário e o computador poderiam interagir acarretando o crescimento das capacidades do usuário. Este sistema surgiu inicialmete como uma ferramenta experimental e mais tarde evoluiu para o NLS/Augment. Já nesta época o sistema utilizava os conceitos de "janela" e "visões" para selecionar informações.

Paralelamente ao desenvolvimento do sistema NLS, TED NELSON (Nelson, 1987b) buscava definir um ambiente unificado de hipertexto capaz de conter toda a literatura mundial com referências interligadas . Este projeto, denominado Xanadu, teve início em 1960 e o primeiro protótipo só esteve disponível para testes "on-line" em 1987.

A partir daí novos sistemas começaram a surgir. Novos encontros e conferências específicas sobre hipertexto e sobre possíveis áreas relacionadas a ele, reunindo profissionais de diversas áreas, têm acontecido com freqüência. Também uma gama imensa de publicações e idéias novas têm surgido. Acredita-se que muito ainda precisa ser feito na área de hipertexto, tanto em termos de implementações, quanto em termos de consolidação de conceitos. Entretanto apesar de algumas dificuldades, ainda não solucionadas, os sistemas de hipertexto começam a se tornar uma realidade, como uma forma séria, flexível e sensível de usar o computador em processos de leitura e escrita.

II.3 - Principais Elementos

A não linearidade na organização da informação em um sistema de hipertexto se relaciona fortemente com a estrutura na qual ele se baseia para associar seus elementos básicos: nós e ligações.

Nós

Os nós são considerados as unidades básicas de informação em um sistema de hipertexto expressando conceitos e idéias tanto do leitor quanto do escritor. Eles podem conter não somente fatos como também regras e operações que auxiliam a navegação em uma rede de hipertexto e a criação automática de novas ligações com outros nós.

Nós costumam ser classificados de diferentes formas em diferentes sistemas, entretanto duas formas básicas aparecem com muita freqüência: nós tipados e não tipados. Qualquer tipo de informação pode estar contida em um nó não tipado, não necessitando de nenhum nome ou descrição para aquele nó. Já os nós tipados possuem rótulos (descritores) capazes de identificar a informação neles contida. Ao limitarmos o estilo de informação contida neles, os nós tipados passam a auxiliar não só na classificação dos nós ou na definição das regras e operações neles contidas como também na busca de uma informação de uma área particular de interesse (Fiderio, 1988), (Lima, 1989).

Os nós também podem ser compostos de forma a representar e manipular grupos de nós como entidades únicas. Através da composição de nós, os sistemas de hipertexto passam a oferecer um mecanismo importante tanto para o auxílio ao processo de autoria múltipla (compartilhada) como para o desenvolvimento paralelo de trabalhos.

No processo de autoria múltipla diferentes autores desenvolvem trabalhos diferentes em contextos diferentes para produzirem um produto final único. Já no desenvolvimento paralelo, diversos módulos de um mesmo trabalho são desenvolvidos em paralelo para uma composição final (Vieira, 1989).

Várias são as abordagens existentes na literatura de hipertexto para esse tipo de nó (composto). O sistema Notecards (Halasz, 1988), por exemplo, utiliza o conceito de "fileboxes" para oferecer ao usuário algumas características do mecanismo de

composição, buscando sobretudo a construção de uma estrutura hierárquica do documento. Os "fileboxes" são nós especializados que organizam ou categorizam vários outros nós ou mesmo outros "fileboxes".

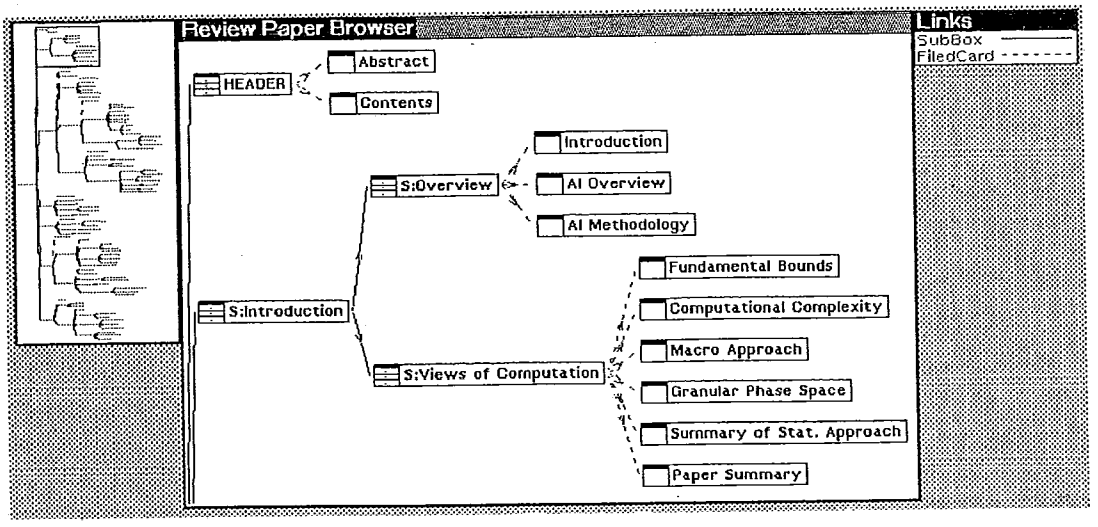


Fig. II.2 - Exemplo de "fileboxes" no Sistema Notecards

Já no Intermedia (Smith & Zdonik, 1987), (Meyrowitz, 1986), contextos diferentes são representados através dos "webs" (ou mapas). Através desses "mapas" o sistema auxilia o usuário a navegar e compreender melhor o universo textual considerado.

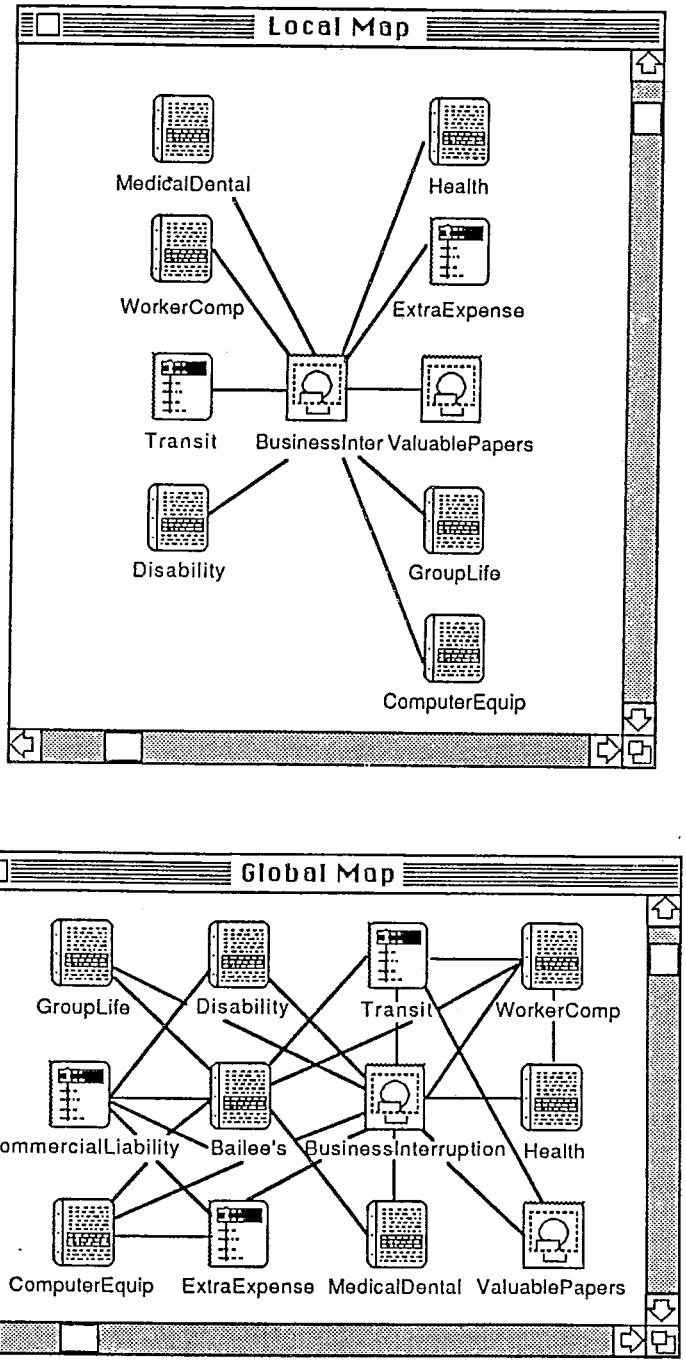


Fig. II.3 - Mapas local e global no Sistema Intermedia

SOARES ET AL (Soares et al, 1990) propõem a utilização de "nós contexto" capazes de estruturar hierarquicamente um documento e definir visões diferentes desse mesmo documento de acordo com a aplicação ou a classe de usuários. Esse tipo de nó agrupa um conjunto de "nós terminais" ou de contexto, recursivamente. Um nó terminal armazena dados (textos, figuras, imagens, etc.) que são apresentados num meio de saída.

À medida que as idéias se desenvolvem e o conhecimento do espaço de informação evolui, organizações pré-definidas vão-se tornando obsoletas. Dessa forma, F. HALASZ (Halasz, 1988) sugere a utilização de estruturas determinadas dinamicamente (ou virtuais) onde os nós são definidos intensionalmente. Ou seja, ao invés de especificar exatamente os seus componentes, apenas a descrição destes componentes é especificada. Assim sendo, nós poderão ser criados dinamicamente (instanciados) no momento em que eles são acessados, possibilitando ao usuário adicionar propriedades ou descrições complementares.

Entretanto instanciar uma estrutura virtual significa satisfazer consultas e construir entidades dinâmicas dos resultados dessas consultas. Logo, a ausência de mecanismos eficientes de consultas e o tempo de resposta envolvido, ainda representam obstáculos à implementação dessas estruturas.

Ligações

As ligações são formas utilizadas para manter os nós conectados, tornando explícito o relacionamento existentes entre eles.

Além de conectar dois nós, as ligações podem também conectar anotações para um documento tais como notas ou comentários, partes sucessivas de textos, e fornecer recursos adicionais conectando gráficos ou outro tipo de informação necessária para tornar a compreensão do texto mais clara e precisa. Enfim, ligações são utilizadas para fornecer uma recuperação seletiva da informação contida na rede de hipertexto.

Muitos sistemas permitem que o usuário crie as ligações necessárias, colocando seus respectivos "labels". Outros produtos têm a capacidade de criá-los automaticamente (uma ferramenta importante principalmente para sistemas que necessitam de referência cruzada para grandes bancos de dados de textos) e fornecem também mecanismos para eliminar, alterar os nomes e os atributos de ligações.

Ligações tipadas auxiliam muito na escolha entre os vários possíveis nós a serem visitados. Ou seja, com base no tipo de ligação que conecta os nós, o leitor poderá escolher o nó mais apropriado.

Outras classificações são feitas na literatura de hipertexto com relação aos tipos de ligações. CONKLIN (Conklin, 1987) considera basicamente dois tipos: ligações referenciais e organizacionais. Ligações referenciais conectam nós a pontos ou

regiões do texto e vice-versa. é o tipo mais encontrado nos sistemas de hipertexto. Ligações organizacionais implementam uma hierarquia na organização da informação, ou seja, formam uma árvore (subgrafo) dentro de uma rede (grafo) de hipertexto.

S. DeROSE (DeRose, 1989) estabelece uma taxonomia mais detalhada, classificando as ligações basicamente como intensionais e extensionais. As intensionais seguem fundamentalmente a estrutura e o conteúdo dos textos ligados, não necessitando de nenhuma estrutura de armazenamento para elas. Ou seja, o destino de uma ligação intensional é definido por alguma função sempre que ela for referenciada. Ao invés de especificar exatamente os seus componentes, apenas a descrição destes componentes é especificada. As extensionais englobam o que CONKLIN (Conklin,1987) considera como "referencial". Assim sendo, o destino ou o nó referido é indicado explicitamente e armazenado individualmente numa estrutura do sistema.

Outras taxonomias podem ser encontradas em (Glushko, 1989), (Halasz, 1988) e (Landow, 1987).

II.4 - Classificação dos Sistemas de Hipertexto

JEFF CONKLIN, em (Conklin, 1987), considera que a classificação dos atuais sistemas de acordo com suas características, não é uma tarefa fácil. Desta forma, ele busca classificá-los de acordo com a área de aplicação, identificando quatro grandes grupos:

1. **Sistemas macro-literários**: sistemas que utilizam tecnologias que suportam uma grande biblioteca "on-line". Nestes sistemas as ligações entre os documentos são suportadas por máquina. Nesse grupo se enquadram: o NLS/Augment, proposto por D. ENGELBART; e o Textnet (Trigg & Weiser, 1986) desenvolvido por TRIGG em sua tese de doutorado na Universidade de Maryland.

2. **Sistemas exploratórios**: sistemas que suportam a exploração de problemas fornecendo ferramentas que apoiam um primeiro pensamento não estruturado do problema para a posterior convergência de idéias. São sistemas utilizados para resolver problemas onde os métodos tradicionais não funcionam adequadamente, necessitando de uma estratégia do tipo "força bruta organizada" para solucioná-los.

Neste grupo podemos citar o sistema gIBIS (Issue Based Information System) (Conklin & Begeman, 1989), idealizado com o objetivo de ser utilizado para a análise de problemas que envolvem trocas de idéias, pontos de vista e questionamento buscando sobretudo um entendimento para o problema.

3. **"Browsing Systems"**: sistemas semelhantes aos macro-literários, porém em escala menor, onde a facilidade de uso e o acesso rápido à informação são fundamentais. Dentre esses sistemas podemos citar: o ZOG/KMS (Akscyn et al., 1988) e o Hiperties (Halasz, 1988).

4. Sistemas de Hipertexto Genéricos: são sistemas de propósito geral projetados para suportar uma gama de aplicações de hipertexto. Como exemplo temos: o Intermedia (Smith & Zdonik, 1987) que está sendo desenvolvido na Universidade de Brown; o Notecards (Halasz, 1988), desenvolvido no Xerox Parc por HALASZ, MORAN e TRIGG; o Hypercard (Conklin,1987); e o Guide (OWL, 1988), desenvolvido na Universidade de Kent.

II.5 - Outras Considerações

Basicamente a evolução dos sistemas de hipertexto pode ser dividida em duas fases.

A geração "original", ou primeira geração, tinha como objetivo principal apenas manipular nós do tipo texto sem se preocupar com a possibilidade da inclusão de informações do tipo áudio ou recursos gráficos, por exemplo.

O surgimento de poderosas estações de trabalho abriu espaço para o desenvolvimento de uma nova geração de sistemas utilizando tecnologias mais avançadas, principalmente relacionadas a recursos gráficos.

Além de colocar em disponibilidade diferentes fontes de informação (internas ou externas ao sistema), tal evolução nos leva a repensar a importância de aspectos como flexibilidade e funcionalidade para a disseminação dos sistemas de informação.

É bem verdade que esta nova geração de sistemas apresenta um enorme potencial de utilização que se estende a cada dia para diferentes áreas de aplicação. Entretanto problemas como aqueles encontrados no processo de autoria e navegação ainda representam um entrave muito sério no desenvolvimento desses sistemas. Ou seja, o usuário ainda necessita de ferramentas e mecanismos que lhe auxiliem não só na criação de nós como também na associação de idéias contidas nos nós durante o processo de autoria.

Não basta apenas inserirmos os nós e estabelecermos as ligações. O usuário também tem que ser capaz de navegar nesse conjunto de informações sem se desorientar ou se sentir confuso.

Posteriormente alguns desses problemas serão tratados com mais detalhes.

Capítulo III

Autoria em Sistemas de Hipertexto

"(...) O autor deveria morrer depois de escrever. Para não perturbar o caminho do texto."

(Humberto Eco, "Pós-Escrito a O Nome da Rosa", 1985)

Este capítulo busca detalhar alguns aspectos importantes no processo de criação de uma rede de hipertexto. Adicionalmente, são feitas algumas considerações relativas a navegação nessa rede.

III.1 - Introdução

Em um certo nível, sistemas de hipertexto constituem sistemas fechados de representação do conhecimento. Um Banco de Dados existe e é estruturado inicialmente para auxiliar o usuário a navegar através da informação nele armazenada. Surge então uma questão importante que é o grau em que um sistema de hipertexto resguarda "imperativos autorais". Ou seja: que mecanismos e privilégios são oferecidos ao usuário no sentido de facilitar a construção/gerência de nós e ligações?

Fundamentalmente o processo de autoria envolve a criação de nós e ligações. Este processo muitas vezes tem um impacto muito forte na validação dessa estrutura como uma representação fiel de novas idéias ou de documentos já existentes que serão convertidos em hipertextos, e também na sua posterior utilização.

III.2 - Fragmentação

No processo de criação de um nó muitas idéias não podem ser segmentadas à primeira vista de uma forma estática. Numa

estrutura interna pré-estabelecida e estática o usuário é colocado numa posição de mero manipulador de "objetos-texto" através de uma rede associativa imposta à priori. Sistemas assim concebidos podem impossibilitar uma compreensão mais profunda do seu propósito, não fornecendo ao usuário uma oportunidade de participar ativamente do processo de autoria (Fischer, 1990). Inicialmente o usuário/autor pode não ter uma compreensão suficiente do conteúdo para estruturar a informação adequadamente.

> "(...) Idéias evoluem à medida que nos familiarizamos com a informação, tornando organizações prévias obsoletas"
>
> (Halasz, 1988).

Na maioria das vezes essa evolução poderá acontecer de uma forma colaborativa entre diferentes elementos de um grupo, que participam do processo com diferentes níveis de compreensão, diferentes hipóteses e experiências.

Na criação de um texto é fundamental a troca de comentários para uma possível inserção de alguns desses comentários no corpo do texto, buscando não só aprimorar e auxiliar a sua recomposição como também melhorar a compreensão do seu verdadeiro significado e, mais importante: intensificar a comunicação entre os elementos de um grupo de indivíduos a que esse texto se destina.

Entretanto esse processo de criação e evolução deve acontecer de uma forma bastante apropriada, sendo buscado não só um

relacionamento de compreensão mútua entre escritores e leitores, mas sobretudo a utilização de mecanismos e estratégias que facilitem o relacionamento de idéias (fragmentos de textos) de uma forma coerente, pois a construção/reconstrução pressupõe a existência de uma estrutura de idéias (no caso, multiseqüencial).

É bem verdade que esta concepção se insere nos atuais sistemas de hipertextos e que muitos trabalhos têm sido produzidos com o intuito de fornecer a esses sistemas ferramentas que motivem os usuários a utilizá-los e supram algumas de suas necessidades, principalmente durante o processo de autoria. Porém um projeto pobre de hipertexto ainda é um problema freqüentemente encontrado. O simples fato de quebrar um texto em fragmentos e ligá-los não assegura que ele cumprirá seus objetivos ou será atrativo.

> "Quando aprendemos, nós o fazemos em grande parte porque percebemos e esperamos existir uma seqüência. (...)Quando tais expectativas são violadas, nós experimentamos um sério sentido de rompimento e deslocamento" (Jaynes, 1989).

Quando os usuários seguem as ligações e encontram textos que não parecem ter uma relação significante e coerente com o texto anterior, de onde partiu a ligação, eles se sentem confusos, desapontados. Assim, bons sistemas de hipertexto dependem de um bom projeto de conteúdo, evitando reações negativas por parte dos leitores.

Outros problemas surgem quando buscamos converter para a forma de hipertexto textos que foram originalmente escritos numa forma puramente seqüencial (linear).

Se o texto original foi produzido há muito tempo atrás torna-se necessário um estudo cuidadoso do texto para que estruturas possam ser produzidas, principalmente se os autores e editores não estão presentes para uma eventual consulta. Como resultado, estruturas e fatos implícitos que não foram compreendidos ou reconhecidos poderão ser perdidos no momento da conversão.

Muitos autores, ao definirem o conteúdo dos nós de uma rede de hipertexto, afirmam que esse mesmo conteúdo deverá ser explícito e bem definido e também que os componentes desse nó devem ser capazes de estabelecer relações com outros nós. É importante considerarmos então que muitas idéias colocadas no texto, com muita propriedade, de uma forma implícita não podem ser colocadas em fragmentos discretos de uma forma explícita, correndo-se sempre o risco de perdermos expressividade. O que torna tais textos interessantes é exatamente tudo aquilo que existe de implícito em cada leitura, onde diferentes leitores estabelecem suas próprias "relações". RAYMOND e TOMPA (Raymond & Tompa, 1988) consideram que se uma estrutura explícita for tão expressiva quanto uma estrutura implícita ganha-se com a conversão; caso contrário, a representação em forma de hipertexto será degradada. Dessa forma, hipertextos podem ser inadequados para alguns tipos de textos.

III.3 – Recuperação e Busca de Dados

Percorrer as informações contidas numa rede de hipertexto, principalmente quando se trata de um número considerável de nós, sem se desorientar ou se sentir confuso, ainda é um dos problemas mais sérios num sistema de hipertexto, interferindo diretamente tanto no processo de leitura/compreensão do texto quanto no processo de autoria.

A cada modificação da rede – por exemplo, mediante a criação de um nó – é necessário que o usuário reexamine todos os outros nós para que ele possa garantir que todas as ligações (ou pelo menos aquelas relativamente mais significativas) foram realizadas. Quando se trata de uma rede pequena (p. ex. 50 a 250 nós) e familiar, usuários normalmente não encontram muita dificuldade na localização da informação. Porém à medida que esta rede evolui, se torna não familiar, e com uma estrutura heterogênea, a navegação passa a ser problemática pois, sem dúvida, além do usuário correr um sério risco de se perder no hiperespaço, ou perder o contexto, uma grande quantidade de material irrelevante pode ser lido desnecessariamente. Em se tratando de um processo de autoria múltipla tais fatos se tornam mais prováveis de acontecer pois, adicionalmente, não sabemos que informação foi inserida na rede. Naturalmente, muitas questões levantadas durante o processo de navegação podem agora se tornar mais complexas. Por exemplo: (i) partindo de um lugar

(nó) específico, quais os possíveis nós que podem ser visitados?; (ii) para onde ir primeiro?; (iii) qual a localização do usuário na rede?; (iv) Como ir para um lugar que sabemos (ou pensamos) que existe?

Muitos sistemas existentes já oferecem ferramentas que de alguma forma se propõem a resolver algumas dessas questões. Dentre elas podemos citar: as ferramentas de recuperação e busca comumente encontradas nos SGBDs (p. ex. busca por cadeia de caracteres, palavras-chaves, ou valor de atributo) e as ferramentas de "browsing". Estas últimas incluem:

(i) os mapas globais, cujo objetivo principal é localizar o leitor em relação ao conjunto de informações contidas na rede. Através de uma visualização da estrutura da rede, suas interligações e os diferentes contextos nelas inseridos, o leitor é capaz de perceber os possíveis caminhos que ele pode percorrer para atingir um determinado nó da rede (Lima, 1989). Sistemas como o Notecards (Halasz, 1988) (Fig. III.1) e o gIBIS (Conklin & Begemen, 1989), por exemplo, permitem que o usuário visualize o conteúdo do "browser" em diferentes níveis de detalhes. Se a rede é muito grande, o nível mais alto de detalhe mostra a estrutura da informação, mas sem nenhuma informação semântica (Fig. III.2).

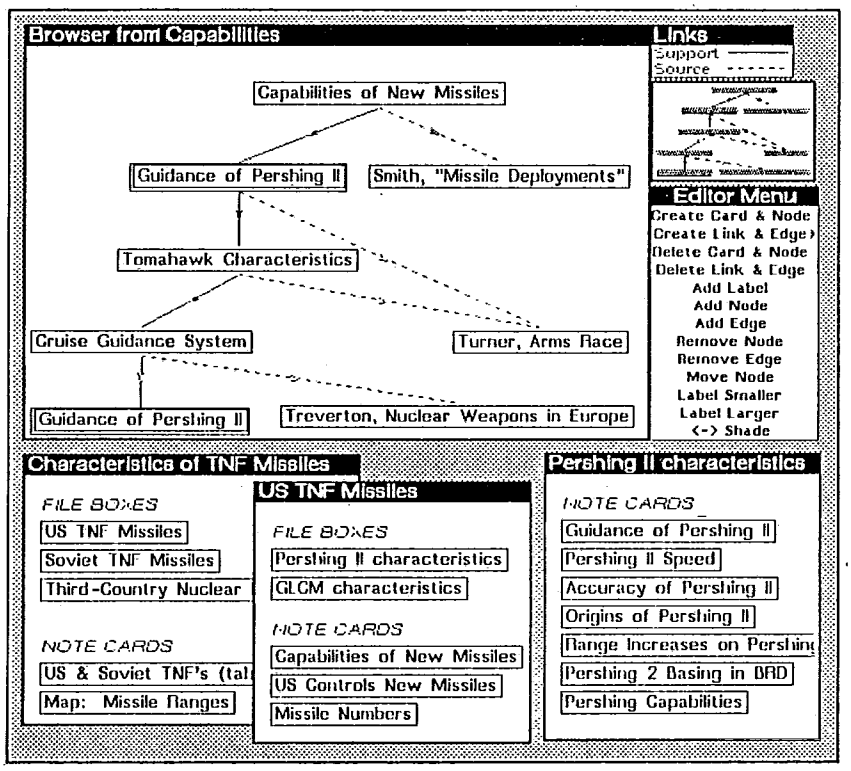


Fig. III.1 - Exemplo de uma tela do Sistema Notecards contendo um mapa global.

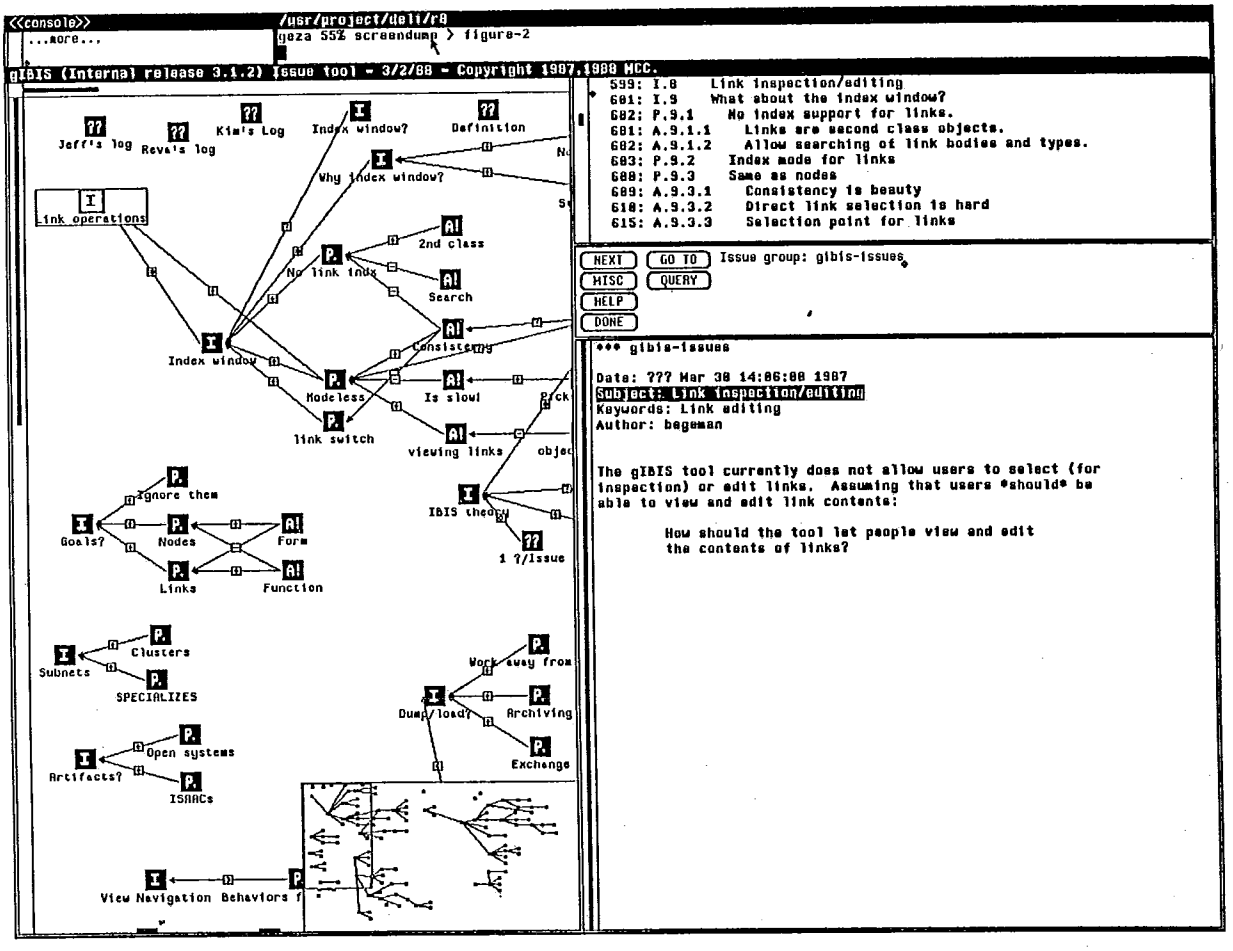


Fig. III.2: Exemplo de um mapa global no sistema gIBIS. No canto inferior, à direita, o sistema mostra uma visão global da rede e a região que está sendo focalizada.

(ii) as listas históricas, que mostram ao usuário todos os nós visitados de uma forma de uma trilha. Alguns sistemas como o Hypercard, por exemplo, apresentam também o tempo gasto durante a visita ao nó (Fig. III.3). Adicionalmente, o sistema Hypercard fornece ao usuário a possibilidade de retornar a um nó, já

visitado anteriormente, através de uma tela histórica gráfica contendo ícones representando os últimos 42 nós visitados. Esta ferramenta auxilia principalmente aquele usuário que apesar de não se lembrar do nome do nó visitado, se lembra perfeitamente do seu aspecto visual.

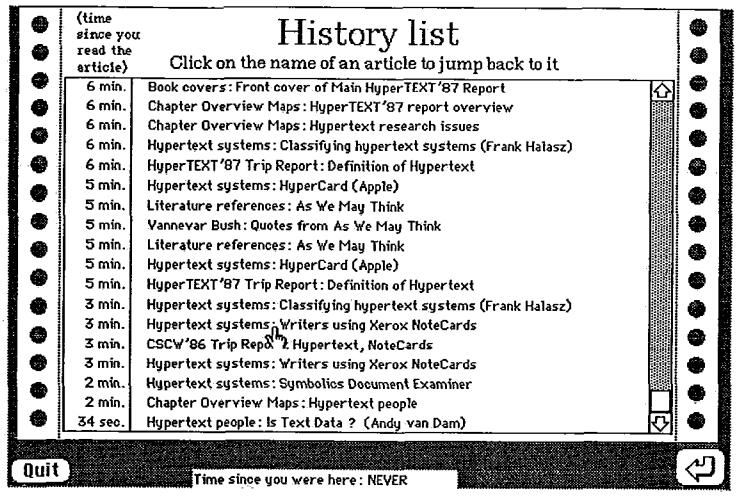


Fig. III.3: Exemplo de uma lista histórica. Ao escolher uma linha da lista, o usuário automáticamente retorna ao nó escolhido.

(iii) as listas históricas gráficas (Notecards) ou os mapas locais (gIBIS), que apresentam ao usuário os nós vizinhos ao nó que está sendo visitado. Assim, pode-se perceber que outros tópicos se relacionam à linha de leitura que está sendo explorada. No sistema Notecards, por exemplo, partindo da lista histórica o usuário poderá selecionar um item (nó) e verificar quais os nós que se conectam àquele selecionado (Fig. III.4).

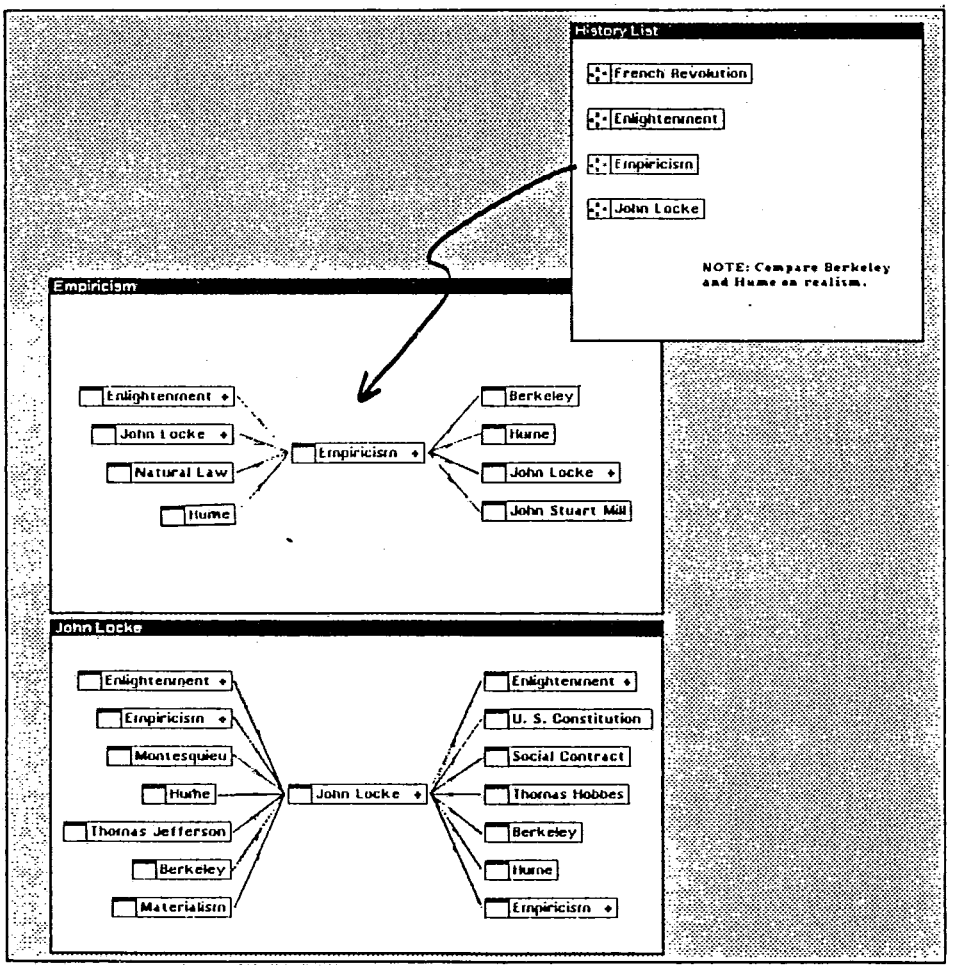


Fig. III.4: Exemplo de uma lista histórica no sistema Notecards

(iv) as árvores históricas (Notecards) que mostram, de uma forma hierárquica, ao usuário que nós foram visitados e como eles foram visitados (Foss, 1987) (FIG. III.5).

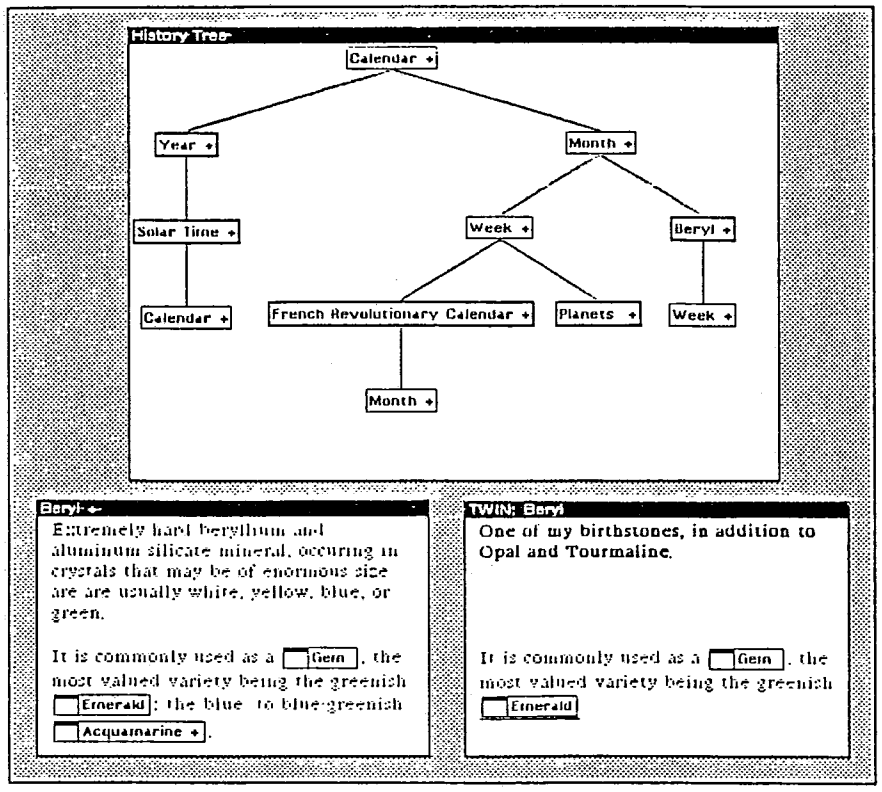


Fig. III.5: Exemplo de uma árvore histórica no sistema Notecards

Alguns sistemas como o Intermedia, o Guide, e o TEXTNET permitem que se estabeleça a priori uma rota ("tours") para guiar ou direcionar o leitor no seu percurso (Fiderio, 1988), (Trigg, 1989). Essas rotas são importantes principalmente na retomada do processo de leitura. Após uma interrupção, muitas vezes o leitor necessita saber qual era a sua posição antes de interromper o processo. Entretanto muitos acreditam que tal solução vem de encontro a uma das características responsáveis pelo poder dos hipertextos/hipermídias que é a liberdade de escolher por onde iniciar, e quais referências percorrer durante uma consulta ao sistema.

Cada leitor pode então selecionar diferentes fragmentos de textos em seqüências diferentes, e para um dado leitor diferentes seqüências podem ser escolhidas em diferentes ocasiões.

Apesar de inúmeros esforços, poucas pesquisas têm sido dirigidas no sentido de verificar até que ponto essa "liberdade" interfere , de alguma forma, no processo normal de leitura e conseqüentemente no de criação (autoria). Será que um controle sobre esse processo, mediante o pré-estabelecimento de caminhos, aprimora o processo de compreensão do texto?

KEARSLEY (Kearsley, 1988) considera que a liberdade de escolha acarreta um nível mais profundo de significado e compreensão da informação. Tal liberdade além de permitir um relacionamento de idéias, e não de simples fatos, provoca no leitor um envolvimento maior e um desejo de ler e interagir mais com o sistema.

Por outro lado, deixar o leitor escolher o que ler, além de não assegurar que uma seqüência de leitura consistente foi devidamente realizada, pode causar distração (muito tempo sendo perdido às vezes em caminhos ineficientes) e uma conseqüente desorientação.

> " (...) liberdade para aprender não é uma condição suficiente para assegurar aprendizado efetivo" (Marchionini, 1988).

Na verdade ainda não existe um consenso em relação às principais soluções até então apresentadas para tornar eficiente o processo de navegação em uma rede de hipertexto. O pressuposto

básico, nesse caso, é chegar a uma solução "aceitável" que não fira a essência dos sistemas de hipertexto e, ao mesmo tempo, não inviabilize a sua disseminação e sobrevivência.

III.4 - Gerência de Nós e Ligações

Dois aspectos importantes ainda não totalmente solucionados prejudicam muito o processo de autoria em hipertextos: o gerenciamento de nós e a construção de ligações.

Várias propostas aparecem na literatura para, se não resolver, pelo menos atenuar esses problemas (gerência de nós e ligações). Dentre elas, SHNEIDERMAN e KEARSLEY (Shneiderman, 1988), (Kearsley, 1988), sugerem que a manutenção de uma lista de todos os nós referenciados ou criados auxiliaria muito o processo de gerência de nós. À medida que o sistema evolui esta lista se tornaria essencial para assegurarmos que as devidas citações (referências) serão feitas de uma forma correta e sem redundâncias (sinônimos, por exemplo, devem ser tratados).

HALASZ (Halasz, 1988) propõe a criação de estruturas virtuais ou determinadas dinamicamente. Ao invés de se definir nós e ligações determinando-se seus componentes exatamente, especifica-se apenas as suas descrições. Por exemplo, os nós de uma sub-rede poderiam ser determinados por uma especificação do seguinte tipo: uma sub-rede contendo todos os nós criados por um usuário "X" durante a última semana. Por outro lado, o destino de uma ligação

poderia ser definido, intensionalmente, da seguinte forma: ligue o nó "X" a um nó contendo a ocorrência de um termo "A" com maior freqüência.

É importante observarmos que além de fornecer um mecanismo de gerenciamento de nós e ligações, estruturas assim definidas evitam construções prematuras, permitindo uma constante manutenção das interligações semânticas face às modificações que ocorrem na rede (Kaplan & Maarek, 1989).

Na verdade, criar uma ligação (indicar seu destino) deveria ser um processo tão simples quanto marcar uma frase ou região que servirão como origem (botão). Entretanto diversos fatores impedem a simplicidade desse processo de forma que a completa automatização seria muita pretensão ou então, certamente, acarretaria outros problemas. Dentre esses fatores dois merecem ser mencionados. Primeiro: ligações às vezes são criadas sem nenhuma justificativa formalizada, sendo realizadas simplesmente com base na própria vontade do usuário. Segundo: a ausência de um "modelo de leitor" (que tipo de leitor irá percorrer as ligações do sistema?) fatalmente levaria a criação automática de ligações interessantes para algumas pessoas, mas óbvias, estranhas ou redundantes para outras (Bernstein, 1990). Adicionalmente, existe um sério risco de ligações excessivas serem criadas, tumultuando a visualização no vídeo, estabelecendo uma sobrecarga cognitiva [1]

---

[1] sobrecarga cognitiva: esforço adicional e concentração necessária para manter várias tarefas ou caminhos ao mesmo tempo.

e conseqüentemente uma total desorientação do usuário.

Buscando atenuar, de certa forma, a sobrecarga de ligações o sistema SmarText (Lotus Dev. Corp.) permite que o usuário estabeleça o número máximo e mínimo de ligações que serão criadas.

Vários pesquisadores têm trabalhado no sentido de estabelecer novas representações de estruturas de ligações em hipertexto que poderiam auxiliar a resolução desses problemas, mas limitar o tipo de ligação (semântica, por exemplo) para muitos, à primeira vista, parece ser uma solução mais adequada.

Independente de limitações impostas pelo mecanismo utilizado é imprescindível que o usuário perceba o conteúdo do sistema de hipertexto que ele utiliza, afastando a dúvida com relação a existência de algum nó de interesse no sistema cuja ausência não pode ser constatada.

Uma Questão da Leitura

Com relação ao segundo fator, colocado anteriormente, que dificulta a automatização do processo de criação de ligações, há um aspecto interessante que merece ser abordado, e que ORLANDI (Orlandi, 1987) coloca com muita propriedade ao definir alguns aspectos das condições de produção de leitura. Segundo a autora, dois elementos unificam o processo de leitura: "o contexto e a relação definida do leitor com a situação". Ao ser especificado

tanto o contexto onde se insere o texto (a sua relação com outros textos) quanto a relação do leitor com a situação que envolve o processo de leitura, naturalmente passa a existir uma distância bem definida, e necessária, que por sua vez selecionará o tipo de leitor capaz de apreender o texto na sua totalidade. Sem uma distância suficiente o leitor não prevê, não antecipa fatos, não interage de forma harmoniosa com o texto. Enfim, sem esses dois elementos que unificam e configuram a leitura, o leitor simplesmente lê palavras e sentenças sem compreendê-las, e conseqüentemente sem perceber o sentido e o objetivo do texto.

É bem verdade que ao criar um texto, o autor automaticamente constitui e delineia o leitor ("leitor virtual"). Entretanto nem sempre esse leitor coincidirá com o leitor real do texto havendo então uma distância entre eles que se estabelece no momento da leitura. À medida que o leitor real se aproxima do leitor virtual é desencadeado um processo crescente de significação. Não havendo uma compreensão suficiente do texto, certamente novas posições merecerão ser consideradas por parte do autor no sentido de alterar o texto ou modificar o leitor, dando a ele condições de produção de uma leitura adequada para que ele eventualmente se aproxime do leitor virtual. É importante ressaltarmos a utilização do termo "produção" no sentido de que na leitura há uma verdadeira interação do autor com o leitor do texto (leitura é produzida).

" Nada consola mais o autor de um romance do que descobrir nele leituras nas quais não pensava e que os leitores lhe sugerem " (Humberto Eco, 1985).

Assim, no relacionamento do leitor com o texto, dependendo da forma como é feita essa interação, da sua completude e da sua intensidade, pode-se, de fato, reconhecer o que o autor quis dizer - paráfrase - , pode-se impor um sentido único ao texto, ou também variar amplamente esse sentido - polissemia - a tal ponto de se buscar um significado que não faz parte da intenção de significação do autor. Surge então um limite difícil de ser definido entre os dois tipos de leitura (a parafrástica e a polissêmica): "aquilo que é o mínimo que o texto "diz" e aquilo que o texto não "diz", considerando-se a intertextualidade, os implícitos em geral, etc." (Orlandi, 1987).

Buscando retratar posições assumidas pelo escritor no momento da criação de um texto, AFFONSO ROMANO (Sant'anna, 1991) considera que "talvez o leitor puro, o leitor que só lê, seja o escritor mais perfeito, o escritor mais feliz, pois é autor gracioso de tudo que lê. Autoriza-se em vários estilos. Escreve em vários gêneros".

Teoria da Leitura?

É possível afirmar que os pressupostos fundamentais relativos a leitura, a uma teoria da leitura, talvez tenham começado a ser redimensionados pela chamada Escola de Frankfurt[1] (Estética da Recepção). A categoria do leitor foi aqui, com esta Escola, inserida e posta em questão, no campo dos estudos literários. Hoje em dia pensa-se na possibilidade de os estudos em Teoria da Literatura talvez caminharem, irreversivelmente para uma teoria da leitura.

No campo dos estudos lingüísticos, a chamada análise do discurso não só incorpora os avanços acima referidos, voltando-se para o leitor, para o texto, para a produção da leitura (ao invés do autor, da frase, do sentido já impresso de antemão) e para a lógica do discurso, como também pôs por terra a antiga distinção que fazia do texto literário, somente ele, lugar-tenente de conotação e pluralidade de sentidos, enquanto todas as outras produções eram vistas como linguagem de segunda, linguagem estereotipada, tipificada. Hoje, já não se pode pensar nos textos literários enquanto cofres de alto saber, que eles "falam" o que

---

[1] posterior a todo o conjunto de influências do pensamento estruturalista dos estudos literários, a E. de Frankfurt, no final dos anos 70 e início dos anos 80, foi bastante lida no Brasil por intelectuais como Luiz C. Lima e Regina Zilberman (Zilberman,1989).

todos os outros discursos talvez tenham silenciado.

A lingüística e a teoria da literatura parecem convergir para o mesmo ponto: a análise do discurso, havendo então a necessidade de se pensar não só nas relações entre elementos internos do texto (unidade, coerência e coesão) como também em tudo que envolve este texto. Faz-se necessário sempre perguntar: quem falou?; para quem falou?; por que falou? ... Isso talvez nos leve a um ponto assustador: a constatação de que no campo da linguagem nada talvez tenha um valor em si. Linguagem é invenção, criação, é da ordem do simbólico.

III.5 - Outras Funções Necessárias no Processo de Autoria

Inúmeras funções precisam ser controladas no processo de autoria em ambientes de hipertexto. Dentre elas podemos citar:

(i) botões em frases são marcados por regiões na tela e precisam ser deslocados quando o texto é modificado.

(ii) restrições e privilégios para a execução de determinadas funções, assim como a abrangência dos comandos envolvidos precisam ser estabelecidos.

(iii) a capacidade de deixar o modo autor, testar novas idéias e retornar ao modo anterior facilmente.

(iv) o tratamento de versões.

(v) o gerenciamento de visões.

(vi) a capacidade para listar nomes dos nós, ligações, sinônimos, listas de exclusão, etc.

Além de outras funções normalmente encontradas nos bons editores de texto.

III.6 - Considerações Finais

Um dos aspectos importantes a serem considerados durante um projeto de hipertexto é o tempo gasto pelo usuário para localizar a informação. SHNEIDERMAN (Shneiderman, 1988) considera este aspecto como um dos atributos-chaves - denominados de "Golden Rules of Hypertext" - para um bom projeto.

Porém uma questão que se relaciona diretamente com esse fato é a determinação do tamanho ótimo de um nó. À medida que reduzimos o tamanho e aumentamos o número de nós na rede, aumentamos também o número de acessos aos nós e conseqüentemente o usuário correrá um risco maior de se perder no "hiperespaço".

Pesquisas recentes evidenciam que usuários preferem nós de menor tamanho a nós maiores. Numa pesquisa recentemente realizada na Universidade de Maryland, criou-se um banco de dados com 46 pequenos artigos (variando de 4 a 83 linhas) e 5 artigos maiores (de 104 a 150 linhas). Aos participantes foram dados 30 minutos para responder uma série de questões utilizando-se esses artigos. Todos aqueles que utilizaram os artigos menores responderam mais questões em um tempo menor (Shneiderman, 1988).

Assim, ao projetarmos uma rede de hipertexto é fundamental considerarmos esse fato pois para lermos um texto (documento) não basta apenas que ele seja coerente, tenha sentido, mas também que se navegue sem se perder ou se confundir.

Capítulo IV

Indexação Automática

"You shall know a word by the company it keeps."

(J. Firth, "A Synopsis of Linguistic Theory 1930-1955", 1957)

Este capítulo apresenta alguns conceitos básicos relativos a área de recuperação da informação. São apresentadas também algumas propriedades relativas a importância de um termo no processo de recuperação de textos.

IV.1 - Introdução

A extração de índices diretamente de textos em linguagem natural e os seus tratamentos de uma forma adequada têm sido objeto de estudo de pesquisadores da área de Recuperação da Informação (RI) durante muito tempo.

Preocupados inicialmente com o gerenciamento de textos relacionados à área de ciências naturais, estudiosos têm buscado formas eficientes para armazenamento e recuperação de dados não estruturados, mais precisamente, de documentos (textos) em linguagem natural.

IV.2 - Função de Indexação

Seja T o conjunto de textos e R o conjunto de possíveis representações desse universo. À função f:T -> R chamamos de Função de Indexação e, para cada documento d E T, f(d) é chamado descritor de d, composto basicamente de índices ou atributos (Maarek & Smadja, 1989); (Maarek et al, 1989).

IV.3 - Efetividade na Recuperação da Informação

IV.3.1 - Validação e Precisão

Dois valores freqüentemente usados para medir a efetividade de um sistema de recuperação da informação são: validação e precisão.

Validação é a "proporção de material relevante realmente recuperado" (Salton, 1985). Ou seja, é um valor que indica se todos os documentos relevantes foram recuperados.

$$\text{validação} = \frac{\text{número de textos relevantes e recuperados}}{\text{número total de textos relevantes}}$$

Precisão é a "proporção de material recuperado suposto ser relevante para as necessidades do usuário" (Salton, 1985). Ou seja, este valor indica se o sistema recupera somente documentos relevantes.

$$\text{precisão} = \frac{\text{número de textos relevantes e recuperados}}{\text{número total de textos recuperados}}$$

Na verdade, podemos considerar o primeiro valor como a probabilidade de que um texto relevante seja recuperado, e o segundo como a probabilidade de que um texto recuperado seja relevante.

Em princípio, buscamos sempre recuperar tudo que desejamos (alta validação) e rejeitar tudo aquilo que não queremos (alta precisão). Entretanto muitos estudos que se baseiam em medidas de validação e precisão necessitam de julgamentos de relevância dos documentos recuperados. Por dependerem fundamentalmente do usuário, tais julgamentos acabam sendo, na maioria das vezes, não confiáveis. Assim, alguns autores acreditam que tais avaliações devam ser abandonadas e, paralelamente, que novas técnicas - menos dependentes da opinião do usuário - devam ser desenvolvidas, com o objetivo de determinar textos relevantes (Frakes & Gandel, 1989).

IV.3.2 - Exaustividade e Especificidade

Duas características principais dos sistemas de recuperação da informação também estão fortemente relacionadas à sua efetividade: **exaustividade** da descrição do documento, e **especificidade** de um termo.

Exaustividade é o grau em que o conteúdo do documento é representado por termos atribuídos a ele. Especificidade é o nível de detalhe (exatidão) em que um dado conceito contido num documento é representado por um determinado termo. Ou seja, dizer que um termo é (ou não) específico significa que ele traz (ou não) consigo um significado preciso e detalhado (Salton, 1985); (Sparck Jones, 1972).

De uma forma geral, exaustividade e especificidade se relacionam de uma forma bem clara com os conceitos de validação e precisão mencionados anteriormente. Aumentar a exaustividade significa que o conteúdo de um documento é representado de uma forma mais abrangente, sendo mais provável a recuperação de documentos relevantes, ou seja, um índice de validação maior. Também, uma especificidade maior de um termo significa que a sua descrição está bastante precisa, sendo menos provável a recuperação de documentos não relevantes, ou seja, um índice de precisão maior será obtido.

É bem verdade que um bom sistema de indexação que busca uma representação mais efetiva do conteúdo dos textos deve se basear também em aspectos lingüísticos, principalmente aqueles que se relacionam à análise semântica. Entretanto estes aspectos ainda apresentam alguns inconvenientes relacionados sobretudo à limitações e dificuldades na aplicação dos métodos de análise lingüística.

Assim sendo, a maioria dos métodos de indexação já implementados e propostos na literatura de recuperação da informação se baseia, fundamentalmente, em técnicas estatísticas. São poucos aqueles que adicionam alguma forma elementar de análise gramatical e/ou semântica.

Atingir níveis suficientes de exaustividade no processo de indexação e também de especificidade de um termo tem sido um objetivo constante dos sistemas atuais. De uma forma geral, é bastante viável relacionarmos exaustividade ao número de termos-

índices atribuídos ao documento - por exemplo, os termos com freqüência mais elevada, considerando-se a utilização de métodos estatísticos. Da mesma forma podemos relacionar especificidade ao número de documentos a que um termo está atribuído. Ou seja, quanto menor o número de documentos ligados a um termo, maior a probabilidade de que este termo seja específico (Salton & Yang, 1973). Dessa forma, especificidade tem sido tratada na literatura como uma função do uso do termo, podendo ser interpretada como uma propriedade estatística.

IV.4 - Sistemas de Indexação Baseados em Palavras-Chaves

Inicialmente a maioria dos sistemas que utilizavam palavras-chaves como unidades de indexação se baseavam em uma lista definida a priori. Nestes sistemas os textos eram simplesmente percorridos e suas palavras eram comparadas com aquelas presentes na lista. Embora estes sistemas sejam úteis, eles podem criar expectativas não realísticas por parte do usuário e um provável desapontamento, recuperando documentos inadequados.

Posteriormente, alguns sistemas passaram a incluir informações adicionais de freqüência e adjacência desses termos. Os sistemas SIRE e o SMART (Salton & Buckley, 1987) são alguns exemplos.

É bem verdade que a freqüência de ocorrência de um termo traz consigo muita informação sobre a importância desse termo no

texto. De fato, estudos têm demonstrado que escritores tendem a repetir palavras importantes à medida que o texto evolui. Elas não ocorrem no texto, ou no universo textual, aleatoriamente, com freqüências iguais. Conseqüentemente, formam-se classes de palavras capazes de serem evidenciadas pela freqüência de ocorrência.

> " A justificativa de medir a significância de uma palavra pela sua freqüência se baseia no fato de que um escritor normalmente repete palavras quando ele melhora ou varia seus argumentos (...)" (Luhn, 1958).

Todavia mesmo utilizando informações de freqüência ou outras técnicas que envolvem até mesmo a participação do usuário, determinando, por exemplo, termos mais relevantes, estes sistemas apresentam alguns inconvenientes. O que se nota com freqüência é a ausência de granularidade, principalmente quando se trata de grandes bancos de dados textuais. Ou seja, palavras-chaves não são suficientes para distinguir documentos de forma eficiente quando eles são numerosos, sem levar em conta que palavras ambíguas surgem com mais freqüência principalmente quando lidamos com grandes universo textuais. Também elas não expressam suficientemente o conteúdo ou o significado do texto. Conseqüentemente, muitos textos indexados a uma determinada palavra são recuperados ao mesmo tempo, apresentando assim uma

baixa precisão, sem aumentar a validação do sistema (todos os documentos necessários não são recuperados).

Muita discussão ainda existe com relação à eficiência dos sistemas de recuperação quando aplicados a domínios restritos. FLASS (Flass, 1985), por exemplo, considera que documentos que exibem um alto grau de "criatividade lingüística", envolvendo diversas áreas, são mais propensos a um índice de validação mais baixo do que textos de um domínio específico. Bancos textuais que apresentam um vocabulário irrestrito e conseqüentemente um domínio semântico também irrestrito necessitam de mecanismos de compreensão do texto auxiliando o processo de indexação.

Por outro lado, BLAIR e MARON (Blair & Maron, 1985) argumentam que é bem verdade que sistemas menos criativos lingüisticamente apresentam um conjunto de palavras limitado para expressar uma determinada idéia. Além disso, vários problemas podem ser tratados a um custo razoável em aplicações que apresentam características lingüísticas particulares relacionadas a domínios específicos. Entretanto a utilização de um mesmo conjunto, mesmo com informações estatísticas adicionais, como possíveis unidades de indexação, deixa de ser uma vantagem a partir do momento em que aumenta a probabilidade de recuperação simultânea de muitos textos [1].

A verdade é que, com os atuais sistemas, usuários ainda encontram dificuldades para conseguir níveis razoáveis de

---

[1] O que os autores chamam com freqüência de "sobrecarga na saída"

validação. Um valor baixo desse índice ainda representa um dos maiores entraves para esses sistemas. Estima-se que as suas taxas médias ficam abaixo de 20% (Maarek et al, 1989).

Validação e Precisão: Como Melhorar?

Várias são as técnicas adicionais propostas na literatura atual visando ao aumento dos índices de validação e precisão dos sistemas de recuperação da informação.

Uma técnica freqüentemente utilizada para aumentar a validação consiste em remover os sufixos das palavras. Termos truncados utilizados como unidades de indexação apresentam um escopo mais abrangente para a identificação de um documento do que palavras completas (Salton, 1985); (Blair & Maron, 1985); (Maarek & Smadja, 1989).

Sinônimos e termos mais gerais obtidos de uma hierarquia previamente estabelecida são também sugeridos com muita freqüência na literatura como formas para se conseguir um índice de validação maior.

Com relação à melhoria na precisão dos sistemas, atribuir pesos aos termos tem fornecido uma valiosa contribuição. Esta técnica se baseia no fato de que alguns termos - por exemplo, os mais freqüentes - são mais importantes do que outros e assim podem servir para uma melhor identificação do texto.

A utilização de listas de exclusão também tem sido sugerida como uma forma de serem eliminadas do texto palavras que carregam

consigo pouca semântica. Estas palavras certamente aparecerão com muita freqüência e serão pouco precisas para representar o conteúdo do texto.

Outro dispositivo que visa aumentar a precisão dos sistemas de recuperação se baseia na utilização de associações de palavras contidas no texto, ao invés de um único termo, como unidade de indexação (Salton, 1985); (Salton & Buckley, 1989); (Maarek & Smadja, 1989). Como estabelecer estas associações e por que utilizá-las é um assunto que será tratado posteriormente.

IV.5 - Importância de um Termo

Diferentes razões nos levam a atribuir um peso a um termo num documento. Uma, mais intuitiva, se relaciona simplesmente à preferência do usuário. Ou seja, o usuário pode estar interessado em documentos que contenham um termo A, ao invés de documentos que contenham B por outras razões que não estão relacionadas diretamente à utilização desses termos no universo de documentos. Por outro lado, modelos probabilísticos podem ser utilizados pelo próprio sistema para atribuir pesos aos termos de tal forma que eles possam representar tanto o comportamento desses termos no conjunto de documentos como também refletir a importância deles para fins de representação do conteúdo dos documentos.

Os estudos iniciais de LUHN (Luhn, 1958), um dos pioneiros da indexação automática, a lei de Zipf (Salton, 1983) e mais

recentemente os trabalhos de CROFT e HARPER (Croft & Harper, 1979), ROBERTSON e SPARCK JONES (Robertson & S. Jones, 1976) e SALTON (Salton, 1990) evidenciam que, ao atribuirmos um valor a um termo, visando sobretudo representar o conteúdo de um texto, dois critérios devem ser levados em conta: (i) termos freqüentes são mais importantes do que aqueles raramente encontrados em um texto; (ii) termos que também aparecem com freqüência em muitos outros documentos do universo textual acabam não sendo tão importantes quanto aqueles que são raramente encontrados na coleção como um todo. Entretanto eles não podem ocorrer com uma freqüência muito baixa, de uma forma insignificante.

Tendo em vista que os termos que ocorrem com freqüência muito alta, ou muito baixa, no universo textual não são bons identificadores de conteúdo, LUHN (Luhn, 1958) afirma então que o "poder de resolução" dos termos extraídos de um texto como unidades de indexação atinge seu valor máximo na região central do intervalo de freqüência desses termos (Fig. IV.1).

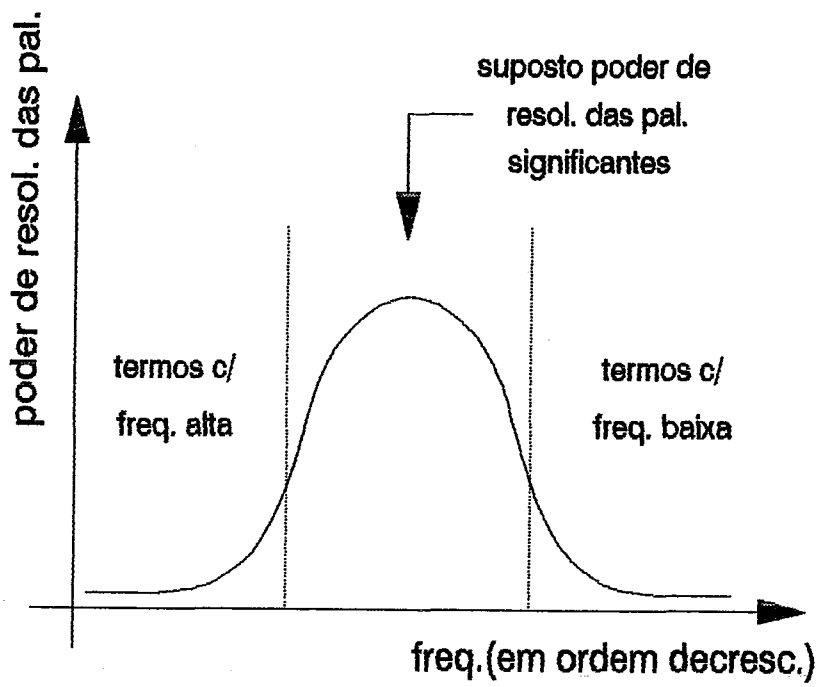


Fig. IV.1 : Poder de Resolução de palavras significantes

Desta forma, SALTON e McGILL definem **poder de resolução** como a "capacidade de um termo identificar itens relevantes e distingui-los dos não relevantes" (Salton & McGill, 1983).

Assim sendo, a importância de um termo, ou o peso atribuído a esse termo como uma possível unidade de indexação se relaciona

diretamente à sua freqüência de ocorrência dentro do documento (freqüência do termo, ft), e à função inversa do número de documentos na coleção em que esse termo aparece como índice (inverso da freqüência de documentos, ifd). Ou seja, uma boa função deve atribuir um peso a uma unidade de indexação que aumenta de valor à medida que sua freqüência local aumenta, mas decresce com a freqüência global no conjunto de textos e também com o tamanho do texto.

Uma possível função que atribua um peso a um termo i que aparece num documento j seria (Salton et al. 1975); (Salton, 1985):

$$w_{ij} = ft_{ij} * ifd$$

O primeiro fator pode ser calculado mediante uma análise do texto considerado. O segundo fator dependerá de uma análise dos textos restantes.

Inicialmente o fator "ifd" foi definido por SPARCK JONES (Sparck Jones, 1972) como:

$$(ifd)_k = (\log_2 n) - (\log_2 d_k) + 1$$

onde d_k é o número de documentos em que o termo k ocorre, num universo de n documentos.

Posteriormente novos estudos propuseram outras formas de se calcular esse valor, buscando sempre atingir índices aceitáveis

de documentos relevantes recuperados (Croft & Harper, 1979), (Salton, 1983), (Salton, 1985).

Na verdade o que se verifica é um forte relacionamento entre a freqüência de um termo no documento e a sua especificidade. SALTON ET ALII (Salton et al, 1975) e SPARCK JONES (Sparck Jones, 1972) sugerem então que termos com alta freqüência, distribuídos uniformemente na coleção como um todo, tendem a ter uma especificidade baixa [1], aqueles com freqüências moderadas tendem a ter uma especificidade moderada e aqueles com baixa freqüência tendem a ter alta especificidade.

Buscando relacionar a especificidade de um termo com o seu ruído que nada mais é do que a medida da "concentração" do termo na coleção, SALTON e McGILL definem o conteúdo de informação de uma palavra w da seguinte forma (Salton & McGill, 1983):

$$info(w) = -\log_2(p_w)$$

onde, p_w é a probabilidade de ocorrência da palavra no universo textual. Esta definição de baseia no fato de que quanto maior a probabilidade de ocorrência de uma palavra menor o seu conteúdo de informação. Por exemplo, se a palavra "hipertexto" ocorre uma vez num universo de 20.000 palavras, sua quantidade de informação será:

---

[1] conseqüentemente, um "ruído alto".

$$info(hipertexto) = -\log_2 5 \times 10^{-5} = 14,29$$

Por outro lado, se ela ocorresse uma vez num universo de 15 palavras, seu conteúdo de informação seria:

$$info(hipertexto) = 3,9$$

IV.6 - Considerações Finais

Trabalhos recentes têm demonstrado que os melhores descritores são aqueles que possuem não só uma distribuição de freqüência moderada na coleção como um todo, mas também uma especificidade moderada.

E mais, termos com uma especificidade não apropriada (baixa ou alta) podem ser tratados no sentido de torná-los adequados como índices. A utilização de sinônimos, por exemplo, poderá diminuir a especificidade. Por outro lado, associações de termos poderão ser usadas para tornar conceitos mais específicos. Este fato será detalhado no próximo capítulo.

Muitos autores acreditam que com o atual estado em que se encontram os trabalhos e as pesquisas sobre o processo de indexação automática, torna-se difícil determinar precisamente o papel da especificidade de um termo nesse processo. Muitos problemas ainda precisam ser solucionados, principalmente aqueles

que se relacionam diretamente com a efetividade do sistema, e refinamentos significantes também precisam ser realizados e testados.

Não basta incorporarmos apenas informações de freqüência dando apenas uma abordagem não sintática. Resultados experimentais recentes evidenciam que termos mais qualificados podem ser determinados se adicionarmos, de alguma forma, informações contextuais ao processo (Fagan, 1989).

Capítulo V

Relações Léxicas

"Tudo o que compõe um estado de língua pode ser reduzido a uma teoria dos sintagmas e a uma teoria das associações. (...)Seria necessário poder reduzir dessa maneira cada fato a sua ordem, sintagmática ou associativa, e coordenar toda a matéria da Gramática sobre esses dois eixos naturais."

(F. de Saussure, "Curso de Lingüística Geral)

Este capítulo apresenta a definição de relações léxicas como unidades de indexação de um texto e como elas podem ser extraídas. Adicionalmente são feitas algumas considerações visando otimizar o processo de extração.

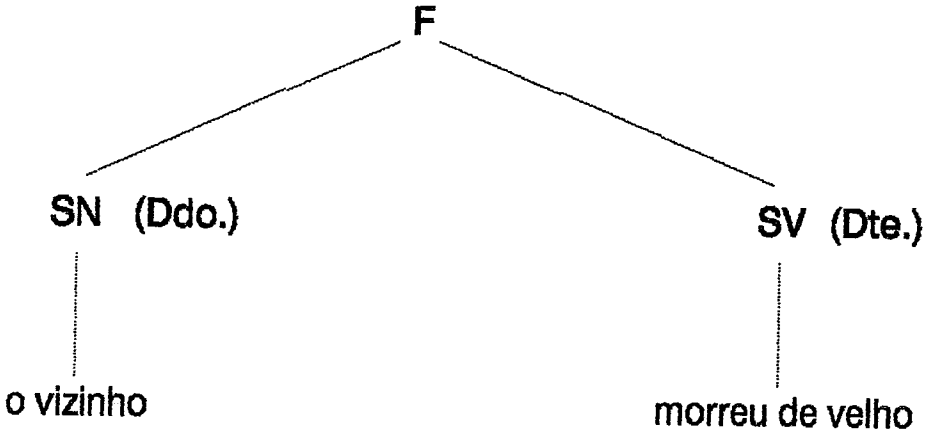
V.1 - Definição

"Relação Léxica entre duas unidades é uma correlação da manifestação comum de dois itens numa sentença" (Maarek & Berry, 1989); (Kaplan & Maarek,1989). Esses itens não ocorrem de forma independente dentro do universo textual, eles são distribuídos de acordo com as regras da linguagem (dependem da estrutura gramatical do texto em que as palavras ocorrem).

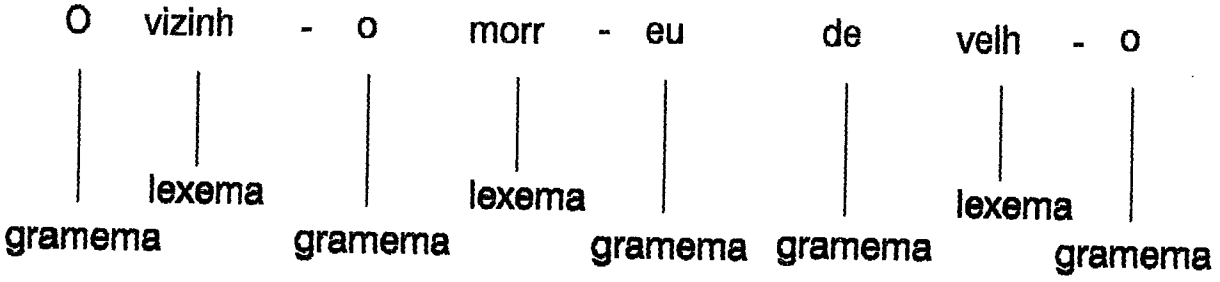
A esse tipo de relação (sintática) ROBINS (Robins,1977) chama de "relação de co-ocorrência". Para formar uma frase ou uma parte especial desta, palavras pertencentes à um determinado grupo ou classe permitem, ou então requerem, a ocorrência de uma outra palavra de um grupo ou classe específica. Assim, um substantivo pode ser precedido por um adjetivo ou artigo. Entretanto um artigo requer a presença de um substantivo ou adjetivo.

Num sentido mais abrangente, à combinação de duas ou mais unidades consecutivas presentes em uma sentença efetiva, onde um termo só adquire seu valor porque se opõe ao que o precede ou ao que o segue SAUSSURE (Saussure, 1949) chama de "relação sintagmática". Ao estabelecermos uma relação desse tipo, passamos a ter um relacionamento de modificação (oposição funcional entre

(i) nível sintático



(ii) nível morfológico



os termos) e também de especificação, onde a relação como um todo passa a fazer referência a conceitos mais específicos [1] .

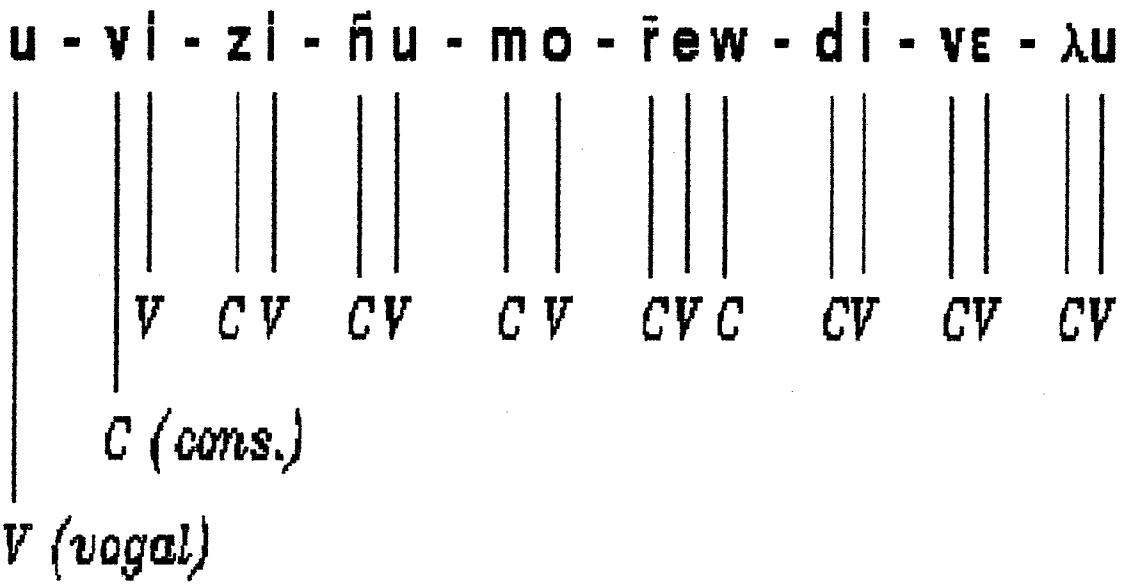
Assim, formas mínimas de significação formam conjugados binários que da mesma forma se combinam dois a dois em novos sintagmas, mais complexos, para expressarem uma determinada idéia.

Cabe ressaltar que a oposição aqui considerada se refere ao nível sintático entre um elemento determinante (sintagma nominal - SN) e um elemento determinado (sintagma verbal - SV), e não ao nível fonológico (onde consoantes e vogais constrastam e instauram o sintagma silábico) ou morfológico (onde lexema e gramema instauram a palavra = sintagma vocabular). Por exemplo, na frase "O vizinho morreu de velho" temos os seguintes níveis:

---

[1] um dos motivos de sua utilização como unidade de indexação.

(iii) nível fonológico



<u>Relações Associativas</u>

Ao definirmos relações léxicas é importante diferenciarmos essas relações daquelas que ocorrem dentro do texto de uma forma associativa. Cada elemento (palavra) ocupa uma posição significativa dentro de um discurso. O significado desse elemento surge não só da posição que ele ocupa referenciando outros elementos que ocorrem ao mesmo tempo no seu contexto, como também através dos elementos exteriores ao discurso (ausentes do contexto considerado) que se associam e formam classes onde imperam diferentes relações. Na verdade, nas relações associativas uma palavra se associa a outras, em um mesmo grupo, por possuírem em comum a mesma "marca semântica" básica. Por exemplo, associado à classe "saúde" aparecem as palavras:

hospital, médico, enfermeira, remédios, etc. Através de uma construção do tipo "palavra-puxa-palavra" esses termos se aproximam, e possuem em comum a mesma marca semântica: "saúde" (Lopes,1975); (Saussure, 1949).

V.2 - Classes de Palavras

Em muitas línguas, palavras se diferenciam formalmente em grupos ou classes seja através de variações paradigmáticas na estrutura morfológica da palavra, ou através dos diferentes tipos de relacionamentos de que elas participam. É bem verdade que, dependendo da complexidade da língua, tais diferenciações ocorrem em maior ou menor grau.

Assim, ao estabelecermos uma classe para uma palavra, formalmente, levamos em consideração não só o seu "comportamento sintático" como também os diferentes "paradigmas morfológicos" que de certa forma vêm reforçar e justificar a sua inclusão numa determinada classe.

Há muito tempo os gramáticos utilizavam nove classes de palavras: substantivo, verbo, pronome, adjetivo, advérbio, preposição, conjunção, artigo e interjeição (Robins, 1977). Entretanto tais classes não foram formalmente definidas de uma maneira clara, apesar de serem justificadas com base em línguas como o latim e o grego, conseqüentemente muitas dificuldades foram encontradas com essa divisão. Os verdadeiros paradigmas

morfológicos e relações sintáticas de que elas participam não foram considerados, acreditando-se apenas que o suposto "conteúdo significativo" da palavra era suficiente para classificá-la.

Várias são as classificações encontradas na literatura. ROBINS (Robins, 1977) e HUDDLESTON (Huddleston, 1984), por exemplo, consideram duas classes de palavras: uma **classe fechada**, que contém um número fixo de palavras e que não se modifica em termos de seus membros, sem uma modificação na estrutura da gramática. Ela engloba os pronomes, as preposições, as conjunções e as interjeições; (ii) uma **classe aberta**, que contém um número ilimitado de palavras que variam de acordo com o falante ou então de acordo com a época. Nela estão incluídos: os substantivos, os adjetivos, os advérbios e os verbos.

Uma das vantagens dessa classificação é que dessa forma evita-se que várias classes sejam estabelecidas de acordo com o comportamento formal da palavra, pois isso acaba criando várias classes para uma mesma palavra. Por exemplo, palavras como "trabalho" pertencem tanto às classes substantivo como verbo.

V.3 - Relações Léxicas: Como Extrair?

Relações Léxicas (sintagmáticas) podem ser construídas de acordo com a relação sintática entre as palavras, ou seja, sujeito-verbo, verbo-objeto direto, etc. Assim sendo, a forma ideal para se extrair tais relações de um texto se daria através

de uma análise sintática. Entretanto a automatização dessa análise certamente nos conduziria a algumas falhas no que diz respeito sobretudo às várias possibilidades que a escrita comporta (estilos não-padrões) (Maarek & Smadja, 1989). Como exemplo podemos citar as diferentes estruturas profundas de uma frase que se associam a uma única estrutura superficial. Para estabelecermos regras que automatizariam o processo certamente teríamos que nos basear nas primeiras.

Consideremos, por exemplo, a frase

O professor trouxe uma lembrança do Rio

ela poderia ter, pelo menos, duas interpretações:

(i) O professor trouxe uma lembrança qualquer, proveniente do Rio

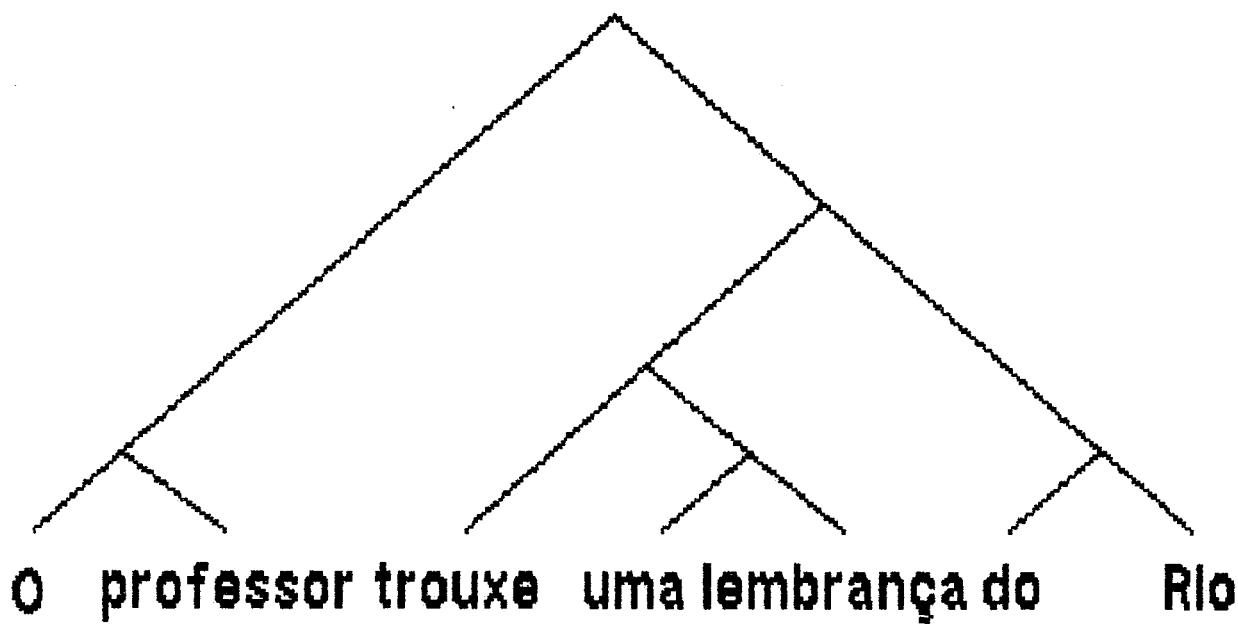


Figura V.3.1 - Representação da Sentença (i)

(ii) O professor trouxe uma lembrança do Rio, proveniente de um lugar qualquer.

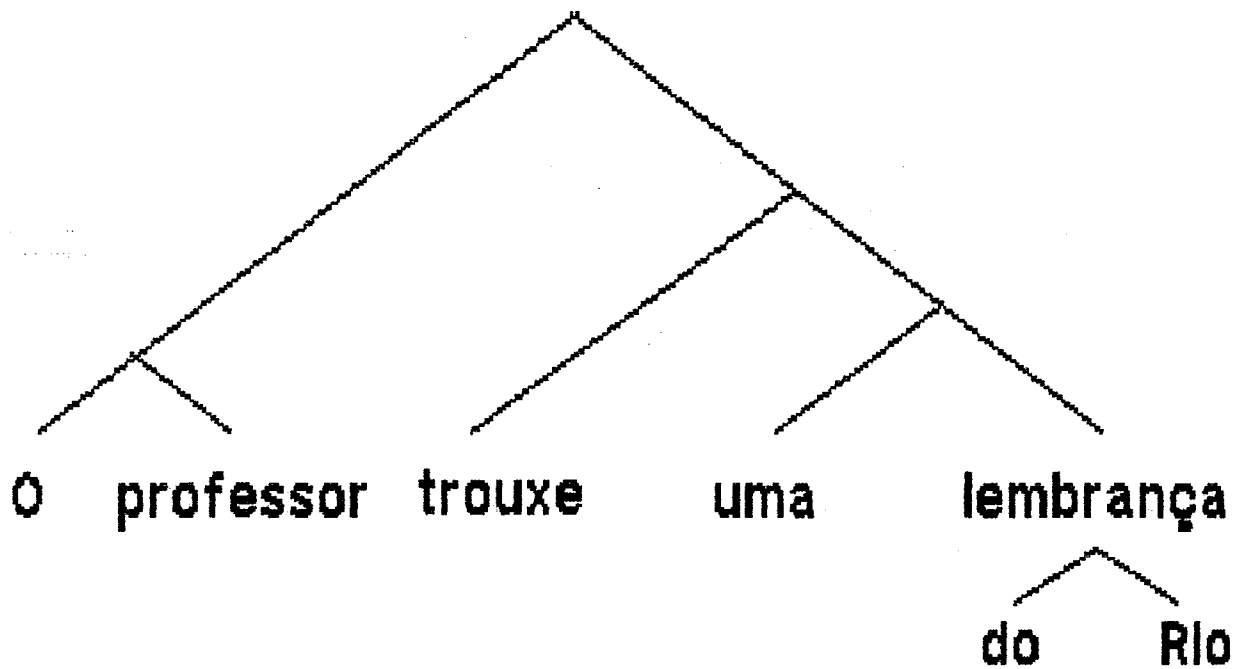


Figura V.3.2 – Representação da Sentença (ii)

Desse modo, uma única estrutura superficial possui, ligadas a ela, duas estruturas profundas (duas diferentes estruturas sintáticas), acarretando duas possíveis interpretações (Chomsky, 1957). Naturalmente este é um assunto que poderia ser tratado mais detalhadamente, porém, tal fato certamente nos levaria a caminhos que fugiriam do escopo do trabalho.

Entretanto isso não é tudo. Mesmo analisando uma frase bem formada, fonológica e sintaticamente (frase gramatical), corremos

o risco de estarmos analisando uma frase sem sentido. Por exemplo: Os óculos viajaram para a lua.

Por outro lado, a gramaticalidade e o sentido de uma frase pode ser ponderada em termos de graus. Certas frases teoricamente agramaticais, em certos contextos, são mais aceitáveis do que outras também agramaticais e sem sentido. Como por exemplo, podemos citar a frase "O povo vieram" que é mais aceitável do que a frase "Vieram povo o". E mais, uma frase como "Os óculos viajaram para a lua" somente parece anormal em certas circunstâncias e contextos, pois se dizemos "é absurdo dizer que os óculos viajaram para a lua" a frase se torna perfeitamente normal, com sentido e gramatical.

A verdade é que existe uma necessidade de contextualizarmos a sentença analisada, adicionando informações provavelmente valiosas que resolveriam alguns impasses.

O consumo de recursos de máquina também precisa ser levado em conta para viabilizarmos a utilização de um analisador (sintático) automático. Estudos têm mostrado que principalmente quando grandes textos são considerados, tais analisadores se tornam inviáveis (Salton, 1989).

Assim sendo, é bastante satisfatória a extração de relações léxicas considerando simplesmente uma vizinhança de uma palavra, adicionando-se a elas um peso que poderá ser, por exemplo, uma função da freqüência e da quantidade de informação da relação no texto considerado. Conseqüentemente aquelas mais importantes dentro do universo textual certamente serão realçadas.

E mais, palavras pertencentes a uma classe fechada (preposições, pronomes, conjunções e interjeições) são fontes de ruído na extração de relações léxicas. Além de serem freqüentes, essas palavras são incapazes de identificar o conteúdo do texto e devem ser eliminadas do processo de construção de relações.

Dessa forma, relações léxicas contendo apenas palavras de uma classe aberta evidenciam uma forma de se colocar mais semântica na descrição de um texto e de fornecer meios para que o contexto seja considerado.

V.3 - Relações Léxicas - Algumas Considerações

Diante do que foi exposto anteriormente, principalmente no que diz respeito à utilização de palavras-chaves como unidades de indexação (ver Cap. IV), podemos afirmar que o principal objetivo da utilização de relações léxicas como unidades de indexação é atingirmos níveis satisfatórios de efetividade no sistema. Entretanto torna-se necessário mencionarmos aqui alguns pontos que se relacionam diretamente com esse fato.

V.3.1 - Lista de Exclusão

Uma lista de exclusão surge da necessidade de filtrarmos as palavras de um texto que podem ser consideradas como termos

"fracos" de indexação. Termos que aparecem com freqüência na maioria dos nós e que realmente não são significantes para representar o conteúdo do texto. Ao serem eliminados do texto, estes índices evitam cálculos supérfluos e otimizam o processo de determinação do conjunto de relações léxicas extraídas do texto.

Assim sendo, esta lista conterá palavras que podem ser consideradas como pertencentes a uma "classe fechada", no contexto considerado.

Tendo em vista a sua estreita relação com o contexto, esta lista necessitará ser revista, de forma um pouco arbitrária, mediante modificações no contexto, pois palavras anteriormente julgadas como necessárias para serem incluídas na lista podem não mais serem.

É bem verdade que a eliminação dos cálculos supérfluos não será perfeita. Para que isto pudesse acontecer, seria necessário uma constante atualização desta lista mediante as freqüentes atualizações dos nós da rede. Esta tarefa se torna muitas vezes inviável.

V.3.2 - Proximidade dos Termos

Ao considerarmos tais descritores (relações léxicas) para representarmos o conteúdo de um texto, um parâmetro essencial deve ser definido: a proximidade dos termos num dado domínio. Este valor, determinado empiricamente, é definido em termos da

distância entre as palavras após serem removidas aquelas que pertencem à classe fechada ou que estão contidas na lista de exclusão [1].

Numa determinada relação léxica que ocorre dentro de uma sentença, quanto mais próximo um termo estiver do outro, maior a probabilidade de relacionamentos significantes serem construídos. Baseados em alguns estudos realizados, alguns autores afirmam que 98% das relações léxicas ocorrem entre palavras que possuem uma distância, no máximo, igual a 5 (Maarek & Smadja, 1989),(Maarek et al, 1989).

V.3.3 - Poder de Resolução

Ao extrairmos de um texto relações léxicas, algumas serão mais importantes do que outras, principalmente quando levamos em consideração o contexto aonde esses termos aparecem. Analisando-se textos especializados em um determinado assunto, alguns termos relevantes e específicos daquele contexto aparecerão com freqüência formando, muitas vezes, relações que não trazem consigo nenhum significado.

Surge então a necessidade de estender o conceito já exposto anteriormente (sobre a importância de um termo dentro de um

---

[1] palavras adjacentes possuem distância igual a 1.

texto) e aplicá-lo ao conjunto de relações obtidas, para que a seleção das relações mais representativas possa ser realizada.

A quantidade de informação de um termo (w) relaciona-se com o peso (importância) atribuído a ele num documento. Ou seja,

$$info(w) = -\log_2 P(w)$$

onde, P(w) é a probabilidade de ocorrência de w no universo textual (Salton, 1983).

Estendendo esse conceito para um par de palavras (w_1, w_2), a quantidade de informação de uma relação léxica poderia ser definida como:

$$info(w_1, w_2) = -\log_2 P(w_1, w_2)$$

onde, da mesma forma, $P(w_1, w_2)$ é a probabilidade de ocorrência do par (w_1, w_2) no universo textual.

Tendo em vista o que foi exposto anteriormente, no que se refere à extração de relações léxicas em um texto, podemos simplificar esse fator considerando os termos dentro do universo textual como variáveis independentes.

É bem verdade que tais termos não ocorrem de uma forma independente em um texto e são distribuídos de forma que regras pré-estabelecidas da linguagem sejam consideradas. Entretanto esta hipótese de independência decorre principalmente das

limitações existentes em relação à automatização de uma análise sintática em um texto (já explicadas anteriormente).

As palavras envolvidas neste processo certamente serão também variáveis dependentes ("puras" relações léxicas) desde que elas tenham uma importância considerável no texto. Conseqüentemente, o método sugerido para extrairmos tais relações, considerando-se apenas a vizinhança de uma palavra pertencente a uma classe aberta, nos permite considerar tal simplificação inicial válida (Maarek & Smadja, 1989).

Diante dessa hipótese podemos considerar a quantidade de informação de uma relação léxica como:

$$info(w_1, w_2) = -\log_2(P(w_1) * P(w_2))$$

O poder de resolução (ρ) de uma relação léxica que aparece em um texto com uma freqüência f, é definido como:

$$\rho(w_1, w_2) = f * info(w_1, w_2)$$

Este fator surge não só da necessidade de filtrarmos pares de termos não significativos como também aqueles pares que não representam realmente uma relação léxica. Por exemplo, ao analisarmos documentos sobre "hipertexto", é bem provável que dentre as relações extraídas apareça algo do tipo: (hipertexto, hipertexto). Isto se deve ao fato de "hipertexto" ser uma palavra freqüente no contexto considerado.

Apesar de freqüentes, tais relações apresentam uma quantidade de informação baixa, conseqüentemente um baixo poder de resolução. Assim sendo, selecionar relações que apresentam valores de ρ relativamente altos nos leva a obter descrições características do texto analisado.

Como já foi visto, formalismos de indexação baseados em relações léxicas permitem a inclusão de mais semântica nas descrições dos documentos, construindo sistemas mais exaustivos do que aqueles que se baseiam unicamente na utilização de palavras-chaves. Entretanto ganhar exaustividade significa um certo risco em perder especificidade. A utilização do conceito mais abrangente de poder de resolução como peso para uma relação certamente nos assegura maior especificidade na descrição do que a simples utilização de freqüência.

V.5 - Considerações Finais

O principal objetivo da utilização de relações léxicas como descritores de um texto é buscar uma melhoria na efetividade do processo de recuperação da informação.

Entretanto diversos fatores relacionados à extração dessas relações influenciam diretamente a sua utilização como unidades de indexação. Dentre elas podemos citar: o domínio de co-ocorrência das palavras e a proximidade dos elementos no domínio considerado (definido em termos da distância entre as palavras).

Possíveis domínios são o documento, o parágrafo, e a sentença. É certo que palavras ocorrendo em um domínio mais restrito formarão relações mais significativas.

Não existem valores pré-estabelecidos para que possamos conseguir bons resultados na recuperação de textos específicos. Para que possamos estabelecer níveis aceitáveis de efetividade na recuperação, esses valores devem ser determinados empiricamente para o universo textual considerado.

Finalmente, não basta apenas identificarmos as relações significativas em um texto. O processo de indexação consiste também em normalizarmos esse descritores que podem diferir na estrutura, mas que se relacionam em termos de significado. Assim, é possível representarmos relações do tipo "recuperação da informação" e "informação recuperada" por um único descritor "recuper inform".

Alguns trabalhos relacionados especificamente com o tratamento de derivações das palavras estão sendo desenvolvidos. Posteriormente eles serão mencionados, como também será detalhado o processo escolhido para a normalização de relações léxicas como unidades de indexação.

Capítulo VI

Sistema de Autoria e Navegação

"Não há bons ventos para quem não sabe aonde quer ir."

(Sabedoria Popular)

Este capítulo tem por objetivo apresentar uma descrição do protótipo implementado e os resultados obtidos durante a fase de testes. Com base nesses resultados são realizadas algumas conclusões visando auxiliar os processos de autoria e navegação.

VI.1 - Considerações Iniciais

Torna-se quase impossível manter uma rede de hipertexto - contendo um número considerável de nós - atualizando ou criando novas ligações mediante freqüentes atualizações dos nós existentes ou criações de novos nós. Tais ligações muitas vezes precisam ser estabelecidas para acomodar novas idéias, soluções, etc. O problema é complexo. Principalmente quando se trata de uma rede com um número considerável de nós, tal tarefa, se realizada manualmente, acaba se tornando impraticável.

A automatização desse processo certamente auxiliaria o usuário no processo de realinhamento da rede à medida que novos testes são introduzidos. Adicionalmente, diminuiria a sobrecarga normalmente imposta ao usuário/autor que, mesmo tendo certeza da existência de um nó, deveria navegar na rede até encontrá-lo, para então estabelecer a ligação. E mais, idéias ou fatos que inicialmente foram considerados sem importância podem agora se tornar imprescindíveis ao processo de produção da leitura, tendo em vista a evolução da rede, e necessitarão ser referenciados (Kaplan & Maarek, 1989).

O ideal é que pudéssemos ter uma ferramenta que automatizasse por completo o processo de criação de ligações, porém tal proposta seria muita pretensão, considerando-se principalmente o fato de que muitas ligações podem ser estabelecidas sem nenhuma justificativa formal.

Assim o propósito específico do sistema implementado é apresentar um mecanismo que possa relacionar conceitos e idéias contidas em um nó, permitindo também que o usuário/autor analise aspectos particulares, e que precisam ser explorados durante o processo de criação de um texto (nó), de produção de uma leitura adequada. Adicionalmente tal mecanismo poderá auxiliar, de certa forma, o processo de navegação na rede.

VI.2 - Descrição Funcional

VI.2.1 - Introdução

O principal objetivo do sistema implementado foi combinar técnicas de Recuperação da Informação e a extração e utilização de relações léxicas que ocorrem de forma significativa no texto, como unidades básicas de indexação. Assim, buscou-se impor ao processo de indexação uma forma de se considerar mais o conteúdo do texto, os conceitos envolvidos, que dificilmente seriam considerados simplesmente através da utilização de palavras-chaves como unidades de indexação.

Tal abordagem visa também atingir níveis aceitáveis tanto de validação quanto de precisão no processo de recuperação de textos.

De posse dos conjuntos de relações extraídas de cada texto e de uma análise desses conjuntos, procurou-se estabelecer uma forma de utilizá-los para auxiliar o processo de autoria em hipertextos, principalmente no que diz respeito a criação/manutenção de ligações.

Inicialmente pensou-se em utilizar tais relações apenas para o estabelecimento de ligações na rede. No entanto com a evolução dos testes e resultados obtidos, novas conclusões foram surgindo em relação à utilização desses índices para auxiliar tanto na construção dos nós como também no processo de navegação na rede.

VI.2.2 - Classe Aberta de Palavras

Termos que participam dessas relações léxicas foram considerados como pertencentes a uma classe aberta de palavras contidas numa sentença (domínio considerado) e separadas por uma distância de, no máximo, cinco palavras.

Para extrair tais palavras foi utilizada uma lista contendo palavras pertencentes a uma classe fechada (interjeições, preposições, artigos, etc.) (Bechara, 1978); (Rocha Lima, 1985) tendo em vista que esta classe contém um número fixo de palavras. Uma vez selecionadas as palavras que participarão de uma relação

léxica dois fatos foram considerados e tratados: extração de sufixos e a utilização de sinônimos.

VI.2.3 - Sufixos

O tratamento de formas variantes das palavras através de algorítmos para extração de sufixos foi, na verdade, um dos primeiros estudos adicionais realizados e implementados nos antigos sistemas de recuperação que se baseavam em palavras-chaves. Atualmente esse mecanismo é tão freqüente nos sistemas de recuperação da informação que a maioria deles naturalmente já não menciona mais a sua utilização, e nem passam a fazer parte da descrição do sistema.

Dentre os trabalhos realizados é importante citar: (i) o mecanismo do sistema SMART (Salton, 1971) que utiliza, respectivamente, uma lista de sufixos, uma lista de exceções e um conjunto de regras para remover alguns sufixos que não puderam ser tratados anteriormente; (ii) a técnica desenvolvida por HAFER e WEISS (Hafer & Weiss, 1974) que utiliza propriedades estatísticas de uma coleção (variação de letras sucessoras e predecessoras) para indicar onde uma palavra deve ser segmentada.

Muitos trabalhos continuam sendo realizados no sentido de produzir algorítmos para que se possa extrair sufixos de uma forma mais eficiente, particularmente em conjuntos de textos específicos ou domínios também específicos (Harman, 1991).

A maioria dos métodos de indexação que se baseiam na análise estatística do texto encontra sérias dificuldades principalmente ao extrair termos significativos e atribuir um peso a eles. Tais métodos acabam tratando a ocorrência das palavras e suas formas variantes (plural, desinências, etc.) como termos independentes. Assim o mesmo conceito, expresso várias vezes de formas diferentes, só será considerado (em termos de ocorrência) se for encontrada, no texto, uma das formas escolhidas para representá-lo. Por exemplo, as palavras "recuperação", "recuperar" e "recuperada" serão tratadas como termos independentes quando, na verdade, poderia ser vantajoso reconhecer o relacionamento semântico entre eles e considerá-los como únicos — p. ex. "recuper" — mediante um processo de normalização léxica. Neste processo os sufixos desses termos são extraídos e apenas as raízes são consideradas como elementos formadores dos índices.

O processo de extração de sufixos utilizado foi implementado e testado tomando-se como base o mecanismo utilizado no sistema SMART (Salton, 1971). Este processo consta basicamente de duas fases. Inicialmente são retirados os sufixos das palavras baseando-se numa lista estabelecida a priori (Bechara, 1978); (Rocha Lima, 1985). Posteriormente algumas regras são verificadas no sentido de tratar alguns casos (exceções) não resolvidos na primeira fase, ou para desfazer algumas operações de extração de sufixos realizadas nessas palavras incorretamente. Dentre essas regras podemos citar: (i) palavras de tamanho igual a dois não foram consideradas nesse processo, pois possivelmente elas nunca

necessitam ser segmentadas e, se consideradas, serviriam apenas para tornar o processo de extração de sufixos mais trabalhoso (Hafer & Weiss, 1974); (ii) um tamanho mínimo de raiz foi estabelecido, no sentido de evitar casos do tipo: palavras como "rindo" teriam a desinência "ndo" retirada, restando apenas "ri" com nenhum significado semântico; (iii) algumas regras estabelecidas anteriormente para retirar o sufixo não se aplicam em determinadas palavras e precisam ser desfeitas. Por exemplo: em "conceito" num processo normal (via lista de sufixos) seria retirado, incorretamnente o "sufixo" ito.

Apesar de não garantir que todas as exceções serão tratadas, foi observado que o mecanismo utilizado muitas vezes reduz palavras de uma mesma família a uma forma comum (podendo ser, ou não, a raiz) produzindo resultados satisfatórios para o sistema. E mais, mesmo tratando algumas palavras que teoricamente poderiam ser consideradas como exceções, erroneamente não altera o resultado global. Por exemplo, em "software" produz-se "softw".

Surge então uma pergunta: como esse processo garantirá a extração daqueles termos mais significativos? Baseando-se no fato de que as pessoas tendem a repetir palavras à medida que variam seus argumentos, então aqueles termos mais significativos certamente terão uma freqüência maior dentro do documento (e relativamente menor no restante) e conseqüentemente um peso maior.

VI.2.4 - Sinônimos

Apesar do processo de extração de sufixos conduzir a uma normalização léxica, possibilitando considerar palavras que possuem as mesmas raízes como únicas, há uma necessidade adicional de relacionarmos semanticamente aquelas que apresentam estruturas diferentes mas são sinônimos.

Este fato precisa ser considerado, principalmente quando se trata de um processo de autoria múltipla pois, em diferentes contextos, com conhecimentos e hábitos lingüísticos também diferentes, usuários escrevem sobre o mesmo fato e utilizam termos diferentes. FURNAS et AL (Furnas et al, 1987) afirmam que o grau em que a utilização de um termo, para descrever um fato, varia, é maior do que o normalmente esperado. Num trabalho realizado recentemente esses autores constataram que duas pessoas escolhem a mesma palavra para descrever um objeto menos do que 20% das vezes.

Também não é uma noção nova que palavras que são similares em significado ocorrem em contextos similares. Ou seja, há um relacionamento entre o grau de similaridade (sinônimo) de um par de palavras e o grau em que seus contextos são similares.

Dicionários têm sido sugerido com freqüência na literatura para tratar sinônimos. Esta sugestão foi testada, e após extrair os sufixos, o protótipo implementado utilizou um dicionário de sinônimos para relacionar palavras e assim, atualizar tanto a tabela de freqüências locais quanto a de freqüências globais.

Tem-se observado que o tratamento de sinônimos e a sua utilização em sistemas de recuperação da informação tem demonstrado um considerável aumento na taxa de validação desses sistemas com um declínio desprezível na precisão (Kristensen & Jarvelin, 1990).

VI.2.5 - Poder de Resolução

Baseando também nos conceitos de exaustividade e especificidade de um termo (vide item IV.3.2), algumas relações do conjunto extraído do texto serão mais significativas do que outras. Classificá-las simplesmente considerando-se a sua freqüência de ocorrência não parece ser uma escolha adequada. Assim, utilizou-se o conceito de poder de resolução (vide item V.3.3) como uma forma de atribuir um peso às relações. Selecionando-se as relações léxicas que possuem um poder de resolução maior ao invés de maior freqüência, reduz-se uma fonte de ruído normalmente encontrado em alguns processos de extração de índices, devido à presença de palavras freqüentes em determinados contextos.

Antes de se calcular o poder de resolução das relações, muitos cálculos supérfluos foram eliminados no sentido de otimizar o processo, pois muitas palavras, apesar de pertencerem a uma classe aberta, não trazem em si muito significado, contribuindo

apenas para tornar o processo mais lento. Para tal, foi criada uma lista de exclusão contendo tais palavras (vide item V.3.1).

VI.3 - Algumas Estruturas de Implementação

Após ter sido extraído o conjunto de relações léxicas locais (dicionário local) ao nó, o sistema armazena essas relações num dicionário (global) de relações léxicas globais. Tal dicionário é necessário pois a cada nó inserido altera-se o dicionário de freqüência global dos termos, e automaticamente modifica-se o poder de resolução das relações dos nós da rede. Provavelmente os conjuntos das relações exportadas (aquelas com maior peso) pelos nós também serão alterados. É através desse dicionário que se estabelece uma tabela de ligações contendo em cada linha o nó origem, a relação léxica e os nós destino. Além de fornecer uma característica dinâmica às ligações ("soft links") (Maarek, 1989) esta tabela será de suma importância para o processo de navegação.

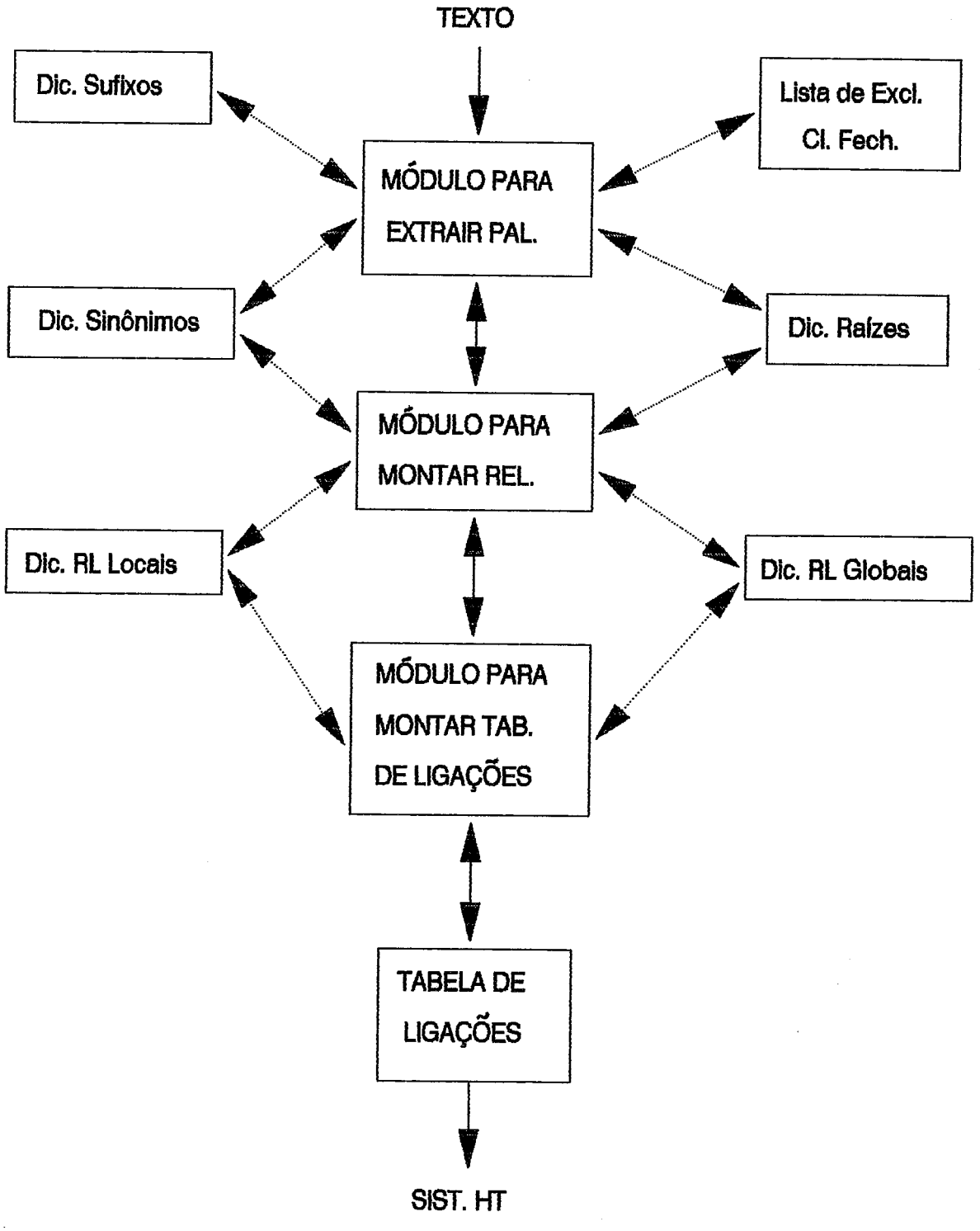


Figura VI.4.1 - Arquitetura do Sistema Implementado

VI.4 - Resultados Obtidos

Ao estabelecermos um conjunto de relações léxicas a serem exportadas por um nó, dois fatos importantes relacionados ao processo de autoria foram observados:

(i) A relação se encontra no conjunto escolhido para ser exportado pelo nó que está sendo inserido.

Neste caso, para que possamos estabelecer uma ligação, devemos buscar nas listas exportadas pelos nós já existentes na rede aquelas que contêm esta relação. Tendo em vista que esta relação pertence tanto à lista do nó inserido como também à dos nós existentes, ela poderá ser um possível botão, bastando apenas definirmos critérios para estabelecer a origem e o destino da ligação.

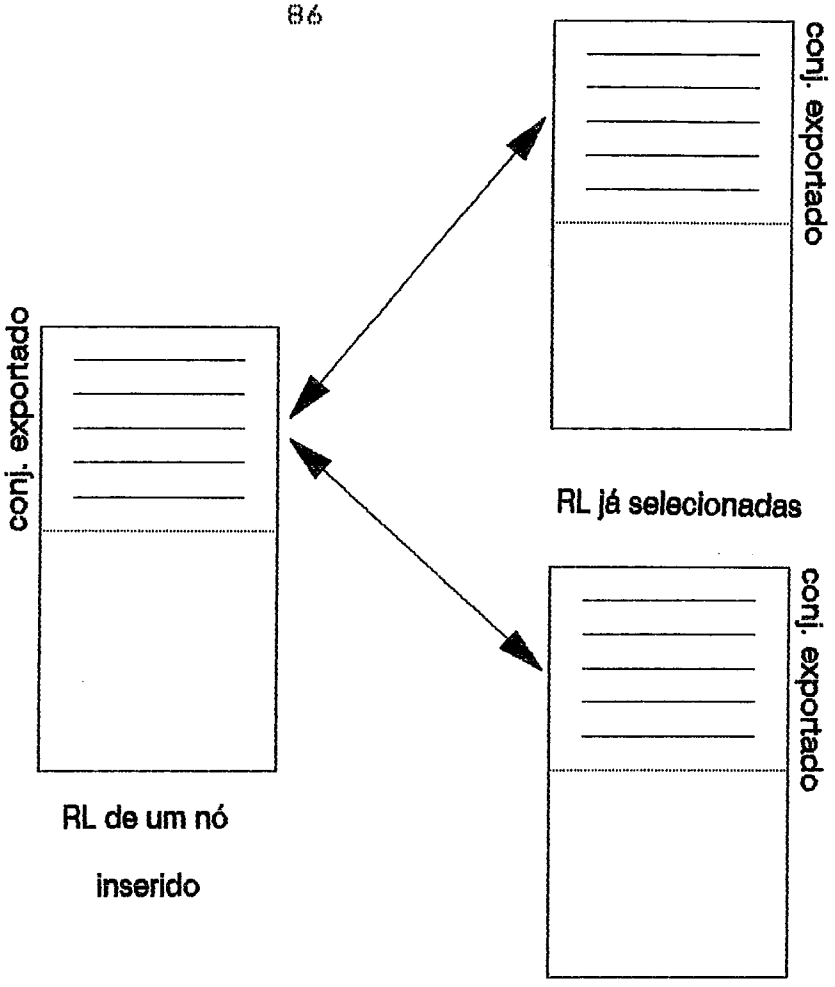


Figura VI.4.2 - Prováveis relacionamentos entre conjuntos de relações léxicas extraídas dos nós.

Entretanto havendo uma interseção dessas listas exportadas, um fato interessante deve ser considerado e analisado: existe um relacionamento semântico entre esses nós estabelecido por esta(s) relação(ções) em comum. A partir desse fato algumas questões devem ser levantadas: será que o nó que está sendo inserido contém a idéia completa? Não está havendo uma fragmentação excessiva dos nós? Ligações desnecessárias não estão sendo criadas?

O principal motivo que leva o usuário/autor a inserir e relacionar um nó em uma rede de hipertexto é a certeza de que

esse mesmo nó adicionará informações valiosas àqueles já existentes formando um "todo maior do que a simples soma das partes". Se dois nós são criados com o mesmo propósito, com o mesmo conteúdo, há necessidade de termos apenas um único nó, ou analisarmos esses nós para verificarmos, por exemplo, se eles deveriam, ou não, estar "acoplados". Tal fato poderá ocorrer com uma freqüência maior principalmente quando estamos num ambiente de autoria múltipla. Um fato - ou uma definição, por exemplo - poderá ser retratado por dois autores de formas diferentes criando-se certamente nós desnecessários.

Ao mencionar alguns aspectos da lingüística textual, Fávero e Koch (Favero & Koch,1983) esclarecem que: " (...) A textualidade - aquilo que faz com que o texto seja um texto - depende, em grande parte de certos fatores responsáveis pela **coesão textual**, conceito semântico que se refere às relações de sentido que se estabelecem entre os enunciados que compõem o texto, fazendo com que a interpretação de um elemento qualquer seja dependente da de outro(s)".

Somente quando os nós diferem de alguma forma é que um deles pode complementar as idéias contidas no outro. Entretanto torna-se difícil estabelecer critérios para diferenciar um texto do outro. Provavelmente esta é uma tarefa que só poderá ser realizada pelo próprio autor. E mais: dependendo da extensão da rede um processo manual poderá ser inviável. Conseqüentemente a informação que sabemos encontrar-se no sistema não será localizada, apesar de estarmos certos de que ela ali se encontra.

Uma análise que leve em consideração as questões expostas anteriormente poderá ser de vital importância principalmente para reduzir a criação de ligações excessivas e desnecessárias. Além de destruir a maioria dos aspectos estruturais e contextuais, tais ligações certamente causariam sérios problemas de desorientação e "sobrecarga cognitiva". E mais, dificultariam tanto a compreensão e manutenção da rede que acabariam excedendo os benefícios trazidos por um formato não linear.

(ii) <u>A relação não se encontra no conjunto exportado</u>.

Com relação a esse fato, dois casos foram observados:

<u>Caso 1</u>: Muitas vezes não sabemos se um determinado nó contendo uma idéia específica se encontra na rede, havendo necessidade, por exemplo, de algum mecanismo de busca para localizá-lo e relacioná-lo com o nó que está sendo editado.

Algumas relações extraídas do nó que está sendo inserido não se encontram na lista exportada pelo nó, no entanto elas poderão ser utilizadas para identificar possíveis nós que se relacionariam com ele. Bastaria para isso que procurássemos nos conjuntos de relações exportadas pelos outros nós aqueles que contêm as relações julgadas importantes no novo nó, que serviriam como possíveis origens (botões) das ligações que partiriam do nó inserido.

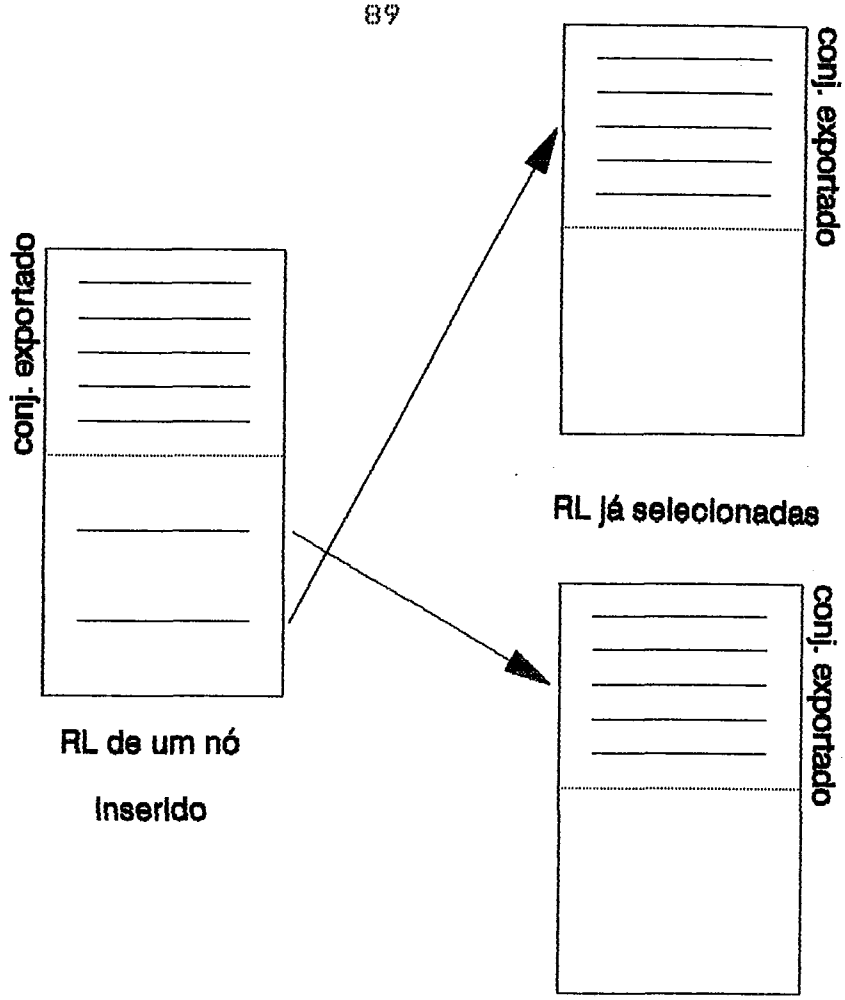


Figura VI.4.3 - Ligações que partem do nó corrente

Assim, ao escolhermos um botão automaticamente o sistema buscaria na tabela de ligações - que por sua vez contém todas as relações exportadas pelos nós -, quais os nós que, provavelmente, serviriam como destino da ligação que partiria daquele botão. Ficaria a critério do usuário/autor confirmar, ou não, o estabelecimento dessa ligação.

Caso 2: Ao inserirmos um novo nó na rede é importante considerarmos também que esse mesmo nó poderia complementar uma idéia de um outro nó, servindo como destino de ligações que partiriam daqueles já existentes na rede.

Esse é um processo um pouco mais complicado pois certamente exigiria que se realizasse uma análise mais criteriosa dos nós já inseridos.

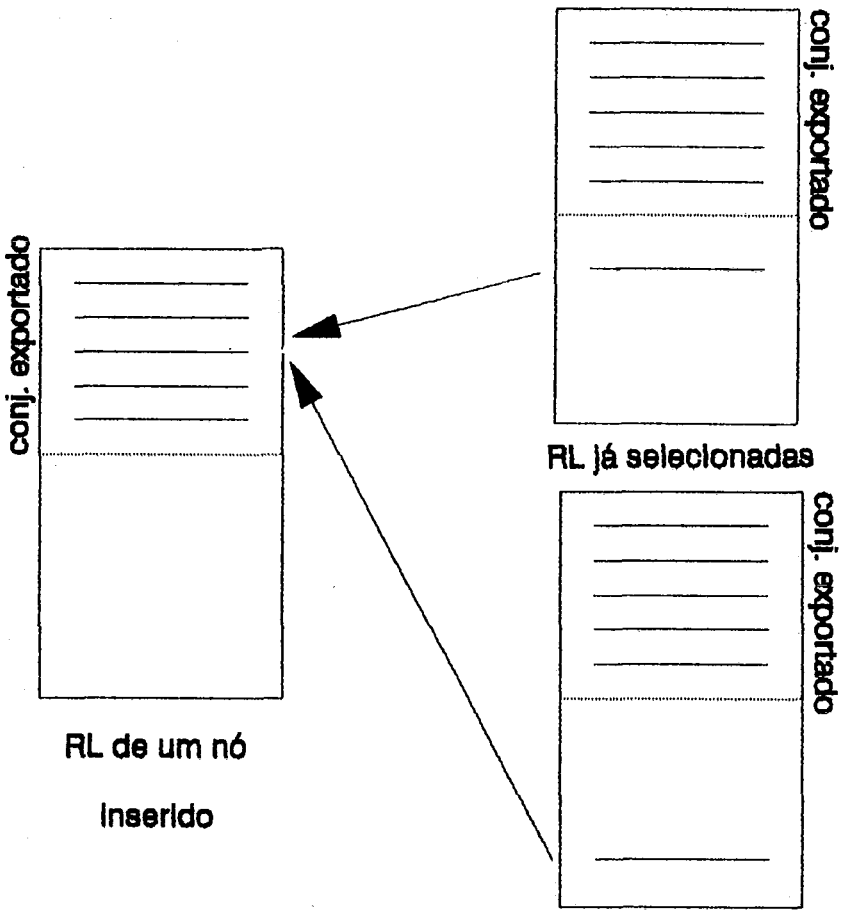


Figura VI.4.4 – Ligações que partem dos nós existentes.

Tanto em relação ao item (i) quanto ao item (ii), os botões podem ser marcados no texto (e, na maioria das vezes são) ou não. Alguns deles muitas vezes são desnecessariamente marcados, podendo ser eliminados e mantidos apenas numa tabela de ligações, evitando dessa forma sobrecarregar visualmente a tela. Dentre esses botões estão as referências explícitas do tipo, "ver cap.", "ver item", etc. que são encontrados com freqüências nos textos.

Hoje questiona-se o fato de termos, ou não, botões marcados explicitamente no texto. Assim, bastaria que o usuário escolhesse um termo e o sistema então realizaria uma consulta a tabela de ligações para verificar a existência, ou não, de uma ligação associada a esse termo.

Navegação

Uma vez estabelecida a tabela de ligações, o processo de navegação seria muito facilitado, bastando para isso consultarmos essa tabela para direcionar o usuário aos prováveis nós destino. Assim, ao escolher uma relação léxica -, marcada como um botão - no texto o sistema mostra ao usuário um menu contendo todos os nós que contém a relação escolhida e foram determinados como destinos das ligações que dali partem.

Na maioria das vezes um botão é uma única palavra, ou então é constituído de mais de duas palavras. Desta forma, há necessidade de se mostrar ao usuário todas as possíveis relações exportadas que contêm aquela palavra e que serviriam para indicar os nós destinos, ou então todas relações que estão contidas no botão escolhido. Uma vez feita a seleção da relação, uma lista dos possíveis nós que a contêm é mostrado ao usuário para que ele prossiga com a navegação.

VI.5 - Considerações Finais

Os processos de extração de sufixos que se baseiam em dicionários e regras são capazes de realizar um processo com alta qualidade. Por outro lado eles ainda dependem muito do esforço humano para construção do dicionário e introdução de regras.

Muitos estudos estão sendo feitos no sentido de facilitar esse processo. Dentre eles é importante citar: (i) o trabalho de HARFER e WEISS (Harfer & Weiss, 1974) que busca definir um processo de segmentação mais automático, sem a necessidade de construções iniciais; (ii) o trabalho de ADANSON e BOREHAN (Adanson & Borehan, 1974) cujo objetivo é desenvolver uma técnica de classificação automática de palavras baseada na sua estrutura de caracteres e na medida do coeficiente de similaridade.

Capítulo VII

Conclusões e Sugestões para Trabalhos Futuros

VII.1 - Considerações sobre o Sistema Proposto

Este trabalho apresentou basicamente uma nova abordagem para amenizar a intervenção do usuário no processo de criação de um hipertexto, levando-se em consideração tanto os aspectos internos (locais) ao texto como também aqueles externos (intertextualidade) que permitirão explicitar as interligações semânticas existentes.

Independente da abordagem ou da ferramenta de autoria proposta, é imprescindível estabelecer uma forma de reduzir não só a "sobrecarga cognitiva" imposta ao autor para lembrar de todos os nós e ligações criadas como também as freqüentes buscas em tabelas contendo índices.

Através de técnicas de recuperação da informação e da utilização de relações léxicas como unidades significativas do texto, a abordagem proposta extrai mais semântica do conteúdo dos nós para o estabelecimento de ligações.

Adicionalmente, buscou-se preservar uma das características mais importantes dos sistemas de hipertextos: a liberdade para criar, para modificar estruturas criadas prematuramente, para detalhar idéias fornecendo ao usuário/autor indicações de nós que possuem um estreito relacionamento semântico entre si. Em muitas situações textos precisam evoluir. Há uma variação histórica em relação a um texto específico, há construções (e leituras) que são possíveis hoje, por exemplo, e que não o foram em outras épocas (Orlandi, 1989).

Com relação a esse ponto, algumas questões importantes relativas à produção da leitura foram levantadas e merecem ser consideradas no processo de criação de um texto.

VII.2 - Conclusões sobre Autoria

O processo de autoria em hipertextos envolve não só a fragmentação da informação em nós como também a especificação (criação) de ligações.

Um conhecimento prévio sobre a estrutura do hipertexto e sobre o seu conteúdo facilita de forma acentuada tanto a fragmentação dos nós como a construção de ligações. Entretanto, nem sempre esse caso acontece. Nem sempre nos deparamos com estruturas familiares, principalmente quando se trata de um sistema de autoria múltipla (co-autoria).

A elevada complexidade cognitiva que envolve tais estruturas certamente dificulta a construção de nós coesos e também o estabelecimento de ligações significativas.

Observa-se freqüentemente que durante o processo de autoria quatro questões naturalmente são levantadas com relação as ligações:

[a] Quais nós, já existentes, devem ser apontados por esse nó?

[b] Quais futuros nós (conhecidos) merecem apontar para esse nó?

[c] Quais nós, já existentes, devem apontar para esse nó?

[d] Quais futuros nós (conhecidos) deveriam ser apontados por esse nó?

Normalmente as questões a e c são respondidas guardando-se um conjunto de todos os nós e ligações, que freqüentemente é varrido e revisto no momento da criação. Esta tarefa se torna exponencialmente complexa à medida que o número de nós e ligações aumenta (Borges, 1991).

Embora não resolvendo por completo tais questões (a, b, c e d), a ferramenta proposta apresentou soluções para a questões "a" e "c" que certamente facilitarão o autor definir explicitamente relações entre textos no momento da criação.

As questões "b" e "d" relacionam-se diretamente a uma construção pré-estabelecida, planejada, de um conjunto de nós. Entretanto, nem sempre esse é o caso. "Escrever é reescrever".

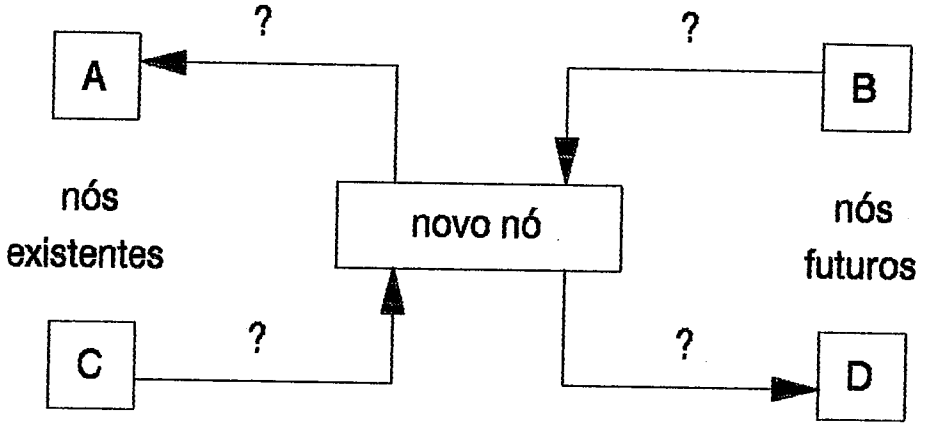


Fig. VII.1: Esquema de ligações em um novo nó

A questão "a" é resolvida apresentando ao usuário/autor uma lista de nós que provavelmente serviriam de nós destinos para as ligações que partem do novo nó.

Também é apresentada uma forma de utilização de relações léxicas para identificar os nós já existentes que possuem um estreito relacionamento, em termos de conceitos neles inseridos, com o nó que está sendo criado. Tal nó merece ser analisado e certificado se a sua existência tem, ou não, uma determinada função na composição da mensagem global.

O mesmo mecanismo também poderá ser utilizado para resolver a questão "c", entretanto os nós existentes necessitarão ser

reanalisados para que o autor possa se certificar da existência de um botão, que servirá como origem de uma ligação, e esta apontará para o nó que está sendo inserido.

Tal análise se refere diretamente ao conjunto de relações exportadas por cada nó, tendo em vista que as ligações aqui consideradas são unidirecionais e normalmente partem de conceitos menos específicos (representados por relações léxicas com baixo poder de resolução) para conceitos mais específicos (representados por relações com alto poder de resolução) em outros nós.

VII.3 - Sugestões para Trabalhos Futuros

Os testes realizados e as experiências adquiridas mostraram como melhorar a qualidade do processo de autoria e como obter índices razoáveis de efetividade. Entretanto, alguns estudos ainda precisam ser realizados, outros caminhos precisam ser explorados para que possamos melhorar tais índices.

É fundamental também ajustar o sistema utilizando-se novos mecanismos em substituição àqueles utilizados e mencionados no capítulo anterior. Como por exemplo:

(i) implementar e validar novas formas para tratar sufixos.

A extração de sufixos poderá reduzir muitas palavras semelhantes morfologicamente a mesma forma, e o efeito é um

agrupamento de palavras baseado em significados similares. Assim, aumenta-se a freqüência atribuída a um termo no processo de indexação trocando-se termos específicos, com freqüência baixa, pelo termo correspondente, mais geral. Entretanto a utilização de sufixos de uma forma irrestrita tem sido freqüentemente evitada pois, certamente, essa forma de agrupamento aumenta o risco de ambigüidade, tendo em vista que raízes trazem em si mais ambigüidade do que palavras completas.

O sistema proposto procurou solucionar, em parte, esse problema, utilizando unidades de indexação maiores. Porém novas formas, que envolvem o tratamento de ambigüidades e a estrutura morfológica das palavras têm sido sugeridas com freqüência na literatura e podem ser encontradas em (Harman, 1991).

(ii) <u>implementar novos mecanismos para gerar e manter listas de exclusão</u>.

Listas de exclusão são utilizadas em sistemas de indexação automática para filtrar palavras que provavelmente seriam índices fracos, e portanto inadequadas para representar o conteúdo do texto.

Freqüentemente estas listas têm incluído somente aqueles termos mais freqüentes num contexto específico, ou palavras cujo conteúdo é tão vago que certamente seriam desnecessárias para o processo de recuperação da informação.

Entretanto, selecionar palavras que irão pertencer a esta lista é uma tarefa mais difícil do que parece, principalmente se

elas são escolhidas com base em dados empíricos a respeito da sua utilização (freqüência). Desta forma, tal lista certamente necessitará de freqüentes revisões e atualizações especialmente quando houver mudança de contexto. Alguns trabalhos que visam testar dentre outras coisas, novas formas de utilização de listas de exclusão podem ser encontrados na literatura, especificamente em (Fox, 1990).

(iii) utilizar um analisador sintático para extrair relações sintagmáticas e validar a sua utilização.

Baseando-se em alguns resultados experimentais FAGAN (Fagan, 1989), afirma que a informação sobre a estrutura do texto, que vai além da simples medida de freqüência e características de co-ocorrência, é necessária. Provavelmente melhores descritores poderão ser construídos, se mais informação sintática puder ser incorporada ao processo de indexação.

SALTON e BUCKLEY (Salton & Buckley, 1989); (Salton, 1989) também consideram que avanços em abordagens sintáticas e semânticas podem, de alguma forma, proporcionar sistemas de indexação mais aceitáveis, com índices de efetividade melhores. Entretanto, com base nos resultados obtidos nos seus estudos comparativos envolvendo sistemas que utilizam uma abordagem sintática, e sistemas que se baseiam somente em características estatísticas de co-ocorrência de palavras, os autores concluem que as abordagens estatísticas ainda apresentam melhores resultados.

Muitas soluções - mais sofisticadas - que envolvem métodos sintáticos têm sido propostas na literatura. O sistema LSP (Linguistic String Project) (Dillon & Gray, 1983), por exemplo, propõe uma análise lingüística mais profunda do texto como a melhor forma de sistemas de indexação mais efetivos serem construídos. Para uma análise sintática mais completa, torna-se necessário um conhecimento semântico maior. Para tal, é utilizado um dicionário de termos.

Da mesma forma, o sistema PHRASE (Earl, 1973); (Dillon & Gray, 1983) utiliza um analisador sintático, porém não incorpora um componente semântico.

O fato é que todas as soluções propostas na literatura evidenciam que dificilmente melhores resultados poderão ser conseguidos, num processo de indexação, sem uma análise sintática e/ou semântica mais profunda do texto, e sem considerações contextuais que possam talvez auxiliar a resolver as ambigüidades porventura existentes.

Outro trabalho que merece ser mencionado é a integração dessa ferramenta com um sistema de hipertexto. Este trabalho já se encontra em andamento no Projeto Hyperbase (Borges, 1991).

A arte de autoria em hipertextos ainda é nova e se encontra em evolução. Espera-se que futuras experiências possam produzir o "feedback" necessário para melhorar os índices de efetividade da ferramenta proposta.

REFERÊNCIAS BIBLIOGRÁFICAS

(AAAI, 1988) AI and Hypertext: Issues and Directions, Proceedings of the AAAI-88 Workshop, EUA, Agosto 1988.

(Adamson & Boreham, 1974) ADAMSON, G., W. e BOREHAM, J., "The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles", Information Storage and Retrieval, Vol.10, No.7, pp.253-260, 1974.

(Akscyn et al., 1988) AKSCYN, R., M., McCRACKEN, D., L. e YODER, E., A., "KMS: A Distributed Hypermedia System for Managing Knowledge in Organization", Communications of the ACM, Vol.31, No.7, pp.820-835, 1988.

(Albuquerque, 1989) ALBUQUERQUE,E.,S.,"O Sistema de Hipertexto H, Dissertação de Mestrado, UFP/Recife, Dezembro 1989.

(Bechara, 1978) BECHARA, E., "Moderna Gramática Portuguesa", 23a. Edição, Companhia Editora Nacional, São Paulo, 1978.

(Bernstein, 1990) BERNSTEIN, M., "An Apprentice that Discovers Hypertext Links", Hypertext: Concepts, Systems and Application, Proceedings of the European Conference on Hypertext, Ed. A. Rizsk, N. Streitz & J. André, pp.212-223, INRIA, France, Novembro 1990.

(Blair & Maron, 1985) BLAIR, D.,C. e MARON, M.,E., "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System", Communications of the ACM, Vol.28, No.3, pp.289-299, 1985.

(Borges, 1991) BORGES, M., R., S.,"Combining Databases and Hypertext: The Hyperbase Project", artigo submetido para publicação, Abril 1991.

(Bush, 1945) BUSH, V., "As We May Thing", Atlantic Montly, pp.101-108, 1945.

(Capodagli, 1989) CAPODAGLI, J., "Hypertext and Hypermedia: Conceptual Design and the Learner", (Fonte Indeterminada), 1989.

(Carr, 1988) CARR, C., "Hypertext: A New Training Tool?", Educational Technology, pp.7-11, Agosto 1988.

(Chomsky, 1957) CHOMSKY, N., Syntactic Structures, Haia: Mouton, 1957.

(Chomsky, 1965) CHOMSKY, N., Aspects of the Theory of Syntax, Cambridge, Mass. : MIT Press, 1965.

(Conklin, 1987) CONKLIN, J., "Hypertext: An Introduction and Survey", IEEE Computer, Vol.20, No.4, pp.17-41, 1987.

(Conklin & Begeman, 1988) CONKLIN, J. e BEGEMAN, M., L., "gIBIS: A Hypertext Tool for Exploratoty Policy Discussion, ACM Transactions on Office Information Systems, Vol.6, No.4, pp.303-331, Outubro 1988.

(Crane & Mylonas, 1988) CRANE, G. e MYLONAS, E., "The Perseus Project: An Interactive Curriculum on Classical Greek Civilization", Educational Technology, pp.25-32, Novembro 1988.

(Croft & Harper, 1979) CROFT, W., B. e HARPER D., J., "Using Probabilistic Models of Document Retrieval without Relevance Information", Journal of Documentation, Vol.35, No.4, pp.285-295, 1979.

(DeRose, 1989) DeROSE, S., "Expanding the Notion of Links", Hypertext'89 Proceedings, Pennsylvania, pp.249-257, novembro 1989.

(Dillon & Gray, 1989) DILLON, M. e GRAY, A., S., "FASIT: A Fully Automatic Syntactically Based Indexing System", Journal of the American Society for Information Science, Vol.40, No.2, pp.115-132, 1989.

(Earl, 1973) EARL, L., L., "Use of word government in resolving syntactic and semantic ambiguities", Information Storage and Retrieval, Vol.9, No.12, pp.639-664, 1973.

(Engelbart, 1963) ENGELBART, D., "A Conceptual Framework for the Augmentation of Man's Intellect", Vistas in Information Handling, P. W. Howerton e D. C. Weeks, Eds., 1963.

(Fagan, 1989) FAGAN, J., L., "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval", Journal of the American Society for Information Science, Vol.40, No.2, pp.115-132, 1989.

(Fávero & Koch, 1983) FÁVERO, L., L. e KOCH, I., G., V., "Lingüística Textual: Introdução", Cortez Editora, 1983.

(Fersko-Weiss, 1991) Fersko-Weiss, H., "3-D Reading with the Hypertext Edge", PC Magazine, pp.241-282, Maio 1991.

(Fiderio, 1988) FIDERIO, J., "A Grand Vision", Byte, Vol.13, No.11, pp.237-244, 1988.

(Fischer, 1990) FISCHER, G., "Communication Requirements for Cooperative Problem Solving Systems", Information Systems, Vol.15, No.1, pp.21-36, 1990.

(Flass, 1985) FLASS, P., R., Technical Correspondence, Communications of the ACM, Vol.28, No.11, pp.1238, 1985.

(Foss, 1987) FOSS, C., L., "Effective Browsing in Hypertext Systems", (Fonte Indeterminada), Novembro 1987.

(Fox, 1990) FOX, C., "A Stop List for General Text", ACM SIGIR FORUM, Vol.24, No.1-2, pp.19-35, 1990.

(Frakes & Gandel, 1989) FRAKES, W., B. e GANDEL, P., B., "Classification Storage and Retrieval of Reusable Components", Proceedings SIGIR'89, 12th International Conference on Research and Development in Information Retrieval, ed. N. J. Belkin e C. J. Rijsbergen, Cambridge, MA, pp.198-206, Junho 1989.

(Furnas et al., 1987) FURNAS, G., W., LANDAUER, T., W., GOMEZ, L., M. e DUMAIS, S., T., "The Vocabulary Problem in Human-System Communication", Communications of the ACM, Vol.30, No.11, pp.964-971, 1987.

(Furuta & Stotts, 1990) FURUTA, R. e STOTTS, P., D., "Generalizing Hypertext: Domains of the Trellis Model", Technique et Science Informatiques, Vol.9, No.6, pp.493-503, 1990.

(Gardin, 1973) GARDIN, J., "Document Analysis and Linguistic Theory", Journal of Documentation, Vol.29, No.2, pp.137-168, 1973.

(Glushko, 1989) GLUSHKO, R., J., "Design Issues for Multi-Document Hypertexts", Hypertext'89 Proceedings, Pennsylvania, pp.51-60, Novembro 1989.

(Goyal, 1989) GOYAL, P., "Intelligent Information Systems: The Concept of an Intelligent Document", Information Systems, Vol.14, No.4, pp.351-358, 1989.

(Greene, 1981) GREENE, J., "Pensamento e Linguagem", 2a. Edição, Zahar Editores, 1981.

(Halasz, 1988) HALASZ, F., G., "Reflections on Notecards: Seven Issues for the Next Generation on Hypermedia Systems", Hypertext'87 Proceedings, North Carolina, pp. 345-366, Novembro 1987.

(Harfer & Weiss, 1974) HARFER, M., A. e WEISS, S., F., "Word Segmentation by Letter Successor Varieties", Information Storage and Retrieval, Vol.10, pp.371-385, 1974.

(Harman, 1991) HARMAN, D., "How Effective Is Suffixing?", Journal of the American Society for Information Science, Vol.42, No.1, pp.7-15, 1991.

(Harris & Cady, 1988) HARRIS, M. e CADY, M., "The Dynamic Process of Creating Hypertext Literature", Educational Technology, pp.33-40, Novembro 1988.

(Huddleston, 1984) HUDDLESTON, R., Introduction to the Grammar of English. Cambridge University Press, 1984.

(Jaynes, 1989) JAYNES, J., T., " Limited Freedom: Linear Reflections on Nonlinear Texts", The Society of Text, Ed. Edward Barret, MIT Press, 1989.

(Kaplan & Maarek, 1989) KAPLAN, S., M. e MAAREK, Y., S., "Incremental Maintenance of Semantic Links in Dinamically Changing Hypertext Systems", Technical Research Report RC15245, IBM Research Division, 1989.

(Kearsley, 1988) KEARSLEY, G., "Authoring Considerations for Hypertext", Educational Technology, pp.21-24, novembro 1988.

(Kritensen & Jarvelin, 1990) KRISTENSEN, J. e JARVELIN, K., "The Effectiveness of a searching thesaurus in free-text searching in a full-text database", International Classification, Vol.17, No.2, pp.77-84,, 1990.

(Landow, 1987) LANDOW, G., P., "Relationally Encoded Links and the Rhetoric of Hypertext", Hypertext'87 Proceedings, North Carolina, pp.331-343, Novembro 1987.

(Lima, 1989) LIMA, M., J., "Hipertexto e suas Aplicações", Projeto de Fim de Curso, Instituto de Matemática, UFRJ, 1989.

(Lopes, 1975) LOPES, E., Fundamentos da Lingüística Contemporânea, 4a. edição, Cultrix, 1975.

(Luhn, 1958) LUHN, M., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol.2, No.2, pp.159-165, 1958.

(Maarek et al., 1989) MAAREK, Y., S., BERRY, D., M. e KAISER, G., E., "Automatically Generating Software Libraries without Pre-Encoded Knowledge", Technical Research Report RC114990, IBM Research Division, 1989.

(Maarek & Berry, 1989) MAAREK, Y., S. e BERRY, D., M., "The Use of Lexical Affinities in Requirements Extraction", Proceedings of the Fifth International Workshop on Software Specification and Design, Pittsburg, PA, pp.196-202, Maio 1989.

(Maarek & Smadja, 1989) MAAREK, Y.,S. e SMADJA, F., A., "Full Text Indexing Based on Lexical Relations - An Application: Software Libraries", Proceedings SIGIR'89, 12th International Conference on Research and Development in Information Retrieval, ed. N. J. Belkin e C. J. Rijsbergen, Cambridge, MA, pp.198-206, Junho 1989.

(Marchionini & Shneiderman, 1988) MARCHIONINI, G. e SHNEIDERMAN, B., "Finding Facts vs. Browsing Knowledge in Hypertext Systems", IEEE Computer, Vol.21, No.1, pp.70-81, Janeiro 1988.

(Nelson, 1987) NELSON, T., Literary Machines, Project Xanadu, 1987.

(Nielsen, 1988) NIELSEN, J., "Trip Report: Hypertext'87", SIGCHI Bulletin, Vol.19, No.4, pp.27-35, 1988.

(Nielsen, 1990) NIELSEN, J., "Hypertext and Hypermedia", Academic Press, Inc., San Diego, CA, 1990.

(Orlandi, 1987) ORLANDI, E., P., "A Linguagem e seu Funcionamento - As Formas do Discurso", Pontes Editores, Campinas, São Paulo, 2a. Edição, 1987.

(OWL, 1988) OWL International Corporation, GUIDE 2: Professional Hypertext System - User Manual, 1988

(Raymond & Tompa, 1988) RAYMOND, D., R. e TOMPA, F., "Hypertext and the Oxford English Dictionary", Communications of the ACM, Vol.31, No.7, pp.871-879, 1988.

(Robins, 1977) ROBINS, R., H., Lingüística Geral, Editora Globo, 1977.

(Robertson & Sparck Jones, 1976) ROBERTSON, S., E. e SPARCK JONES, K., "Relevance Weighting of Search Terms", Journal of the American Society for Information Science, Vol.27, No.3, pp.129-146, 1976.

(Rocha Lima, 1985) ROCHA LIMA, C., H., "Gramática Normativa da Língua Portuguesa", 25a. Edição, José Olympio Editora, 1985.

(Salton, 1973) SALTON, G., "Recent Studies in Automatic Text Analysis and Document Retrieval", Journal of the ACM, Vol.20, No.2, pp.258-278, 1973.

(Salton, 1981) SALTON, G., "A Blueprint for Automatic Indexing", ACM SIGIR Forum, Vol.16, No.2, pp.22-38, 1981.

(Salton, 1985) SALTON, G., "Another Look at Automatic Text Retrieval Sys-tems", Technical Report TR 85-713, Department of Computer Science, Cornell University, Ithaca, N.Y., Dezembro 1985.

(Salton, 1989) SALTON, G., "A Comparison of Book Indexing Methods",Technical Report TR 89-1033, Department of Computer Science, Cornell University, Ithaca, N.Y., Agosto 1989.

(Salton et al., 1975) SALTON, G., YANG, C., S. e YU, C., T., "A Theory of Term Importance in Automatic Text Analysis", Journal of the American Society for Information Science, Janeiro/Fevereiro, pp.33-34, 1975.

(Salton & Buckley, 1987) SALTON, G. e BUCKLEY, C., "Parallel Text Search Methods", Technical Report TR 87-828, Department of Computer Science, Cornell University, Ithaca, N.Y., Abril 1987.

(Salton & Buckley, 1989) SALTON, G. e BUCKLEY, C., "A Comparison Between Statistically and Syntatically Generated Term Phrases, Technical Report TR 89-102Z, Department of Computer Science, Cornell University, Ithaca, N.Y., Julho 1989.

(Salton & Buckley, 1990a) SALTON, G. e BUCKLEY, C., "Improving Retrieval Performance by Relevance Feedback", Journal of the American Society for Information Science, Vol.41, No.4, pp.288-297, 1990.

(Salton & Buckley, 1990b) SALTON, G. e BUCKLEY, C., "Approaches to Text Retrieval for Structured Documents", Technical Report TR 90-1083, Department of Computer Science, Cornell University, Ithaca, N.Y., Janeiro 1990.

(Salton & McGill, 1983) SALTON, G. e McGILL, M., J., Introduction to Modern Information Retrieval, McGraw Hill Computer Series, N.Y., 1983.

(Salton & Yang, 1973) SALTON, G. e YANG, C.,S., "On the Specification of Term Values in Automatic Indexing", Journal of Documentation, Vol.29, No.4, pp.351-372, 1973.

(Sampaio, 1990) SAMPAIO, F., F., "TH - Proposta de uma Ferramenta Automática para Transformação de Textos em Hipertextos", Dissertação de Mestrado, UFRJ/COPPE, Abril 1990.

(Sant'anna, 1991) ROMANO de SANT'ANNA, A., "O Escritor, o Leitor", Jornal O Globo, 17 de Março de 1991.

(Saussure, 1969) SAUSSURE, F., Curso de Linguística Geral, São Paulo, Cultrix/USP, 1969.

(Shasha, 1987) SHASHA, D., "When does Non-Linear Text Help?", Expert Database Systems, Larry Kerschberg, Ed., 1987.

(Shneiderman, 1987) SHNEIDERMAN, B., "User Interface Design for Hyperties Electronic Encyclopedia", Hypertext'87 Proceedings, North Carolina, pp.189-194, Novembro 1987.

(Shneiderman, 1988) SHNEIDERMAN, B., "Reflections on Authoring, Editing and Managing Hypertext", The Society of Text, Vol.9, No.4, pp.34-45, outono 1988.

(Smith & Zdonik, 1987) SMITH, K., E. e ZDONIK, S., B., "Intermedia: A Case Study of the Differences Between Relation Relational and Object-Oriented Database Systems", Proceedings OOPSLA'87, Outubro 1987.

(Soares et al., 1990) SOARES, L., F., G., RODRIGUES, N., RANGEL., J., L., LIMA, M., J., D., TUCHERMAN, L., CASANOVA, M., A., "Delineamento da Arquitetura de um Sistema para Processamento de Hiperdocumentos Multimídia, PUC/RJ, 1990.

(Sparck Jones, 1972) SPARCK JONES, K.. "A Statistical Interpretation of Term Specificity and its Application in Retrieval", Journal of Documentation, Vol.28, No.1, pp.11-21, 1972.

(Sparck Jones, 1974) SPARCK JONES, K.. "Progress in Documentation - Automatic Indexing", Journal of Documentation, Vol.30, No.4, pp.393-432, 1974.

(Trigg, 1989) TRIGG, R., H., "Guided Tours and Tabletops: Tools for Communicating in a Hypertext Environment", ACM Transactions on Office Information Systems, Vol.6, No.4, pp.398-414, Outubro 1988.

(Trigg & Weiser, 1986) TRIGG, R., H. e WEISER, M., "Textnet: A Network-Based Approach to Text Handling", ACM Transactions on Office Information Systems, Vol.4, No.1, pp.1-23, Janeiro 1986.

(Vieira de Melo, 1989) VIEIRA DE MELO, A., C., "Uma Especificação Formal de Links e Nós em um Sistema de Hipertexto para Desenvolvimento de Software, Dissertação de Mestrado, UFP/Recife, 1989.

(Yankelovich & Maygrowitz, 1985) YANKELOVICH, N. e MEYROWITZ, N., "Reading and Writing the Electronic Book", IEEE Computer, Vol.18, No.10, pp.15-29, Outubro 1985.

(Zilberman, 1989) ZILBERMAN, R., "Estética da Recepção e História da Literatura", Editora Ática, São Paulo, 1989.

Apêndice A

Autoria: Um Exemplo

Consideremos uma rede de hipertexto contendo nós que se referem, de uma forma geral, a estrutura e aos serviços prestados pelos diferentes órgãos e departamentos do NCE/UFRJ. Suponhamos também que ao inserirmos o nó denominado "NCE" (Fig. A1) desejamos saber quais os nós (inseridos) que poderiam ser apontados, por exemplo, pelo botão "Helpdesk" (caso a - Cap.VII).

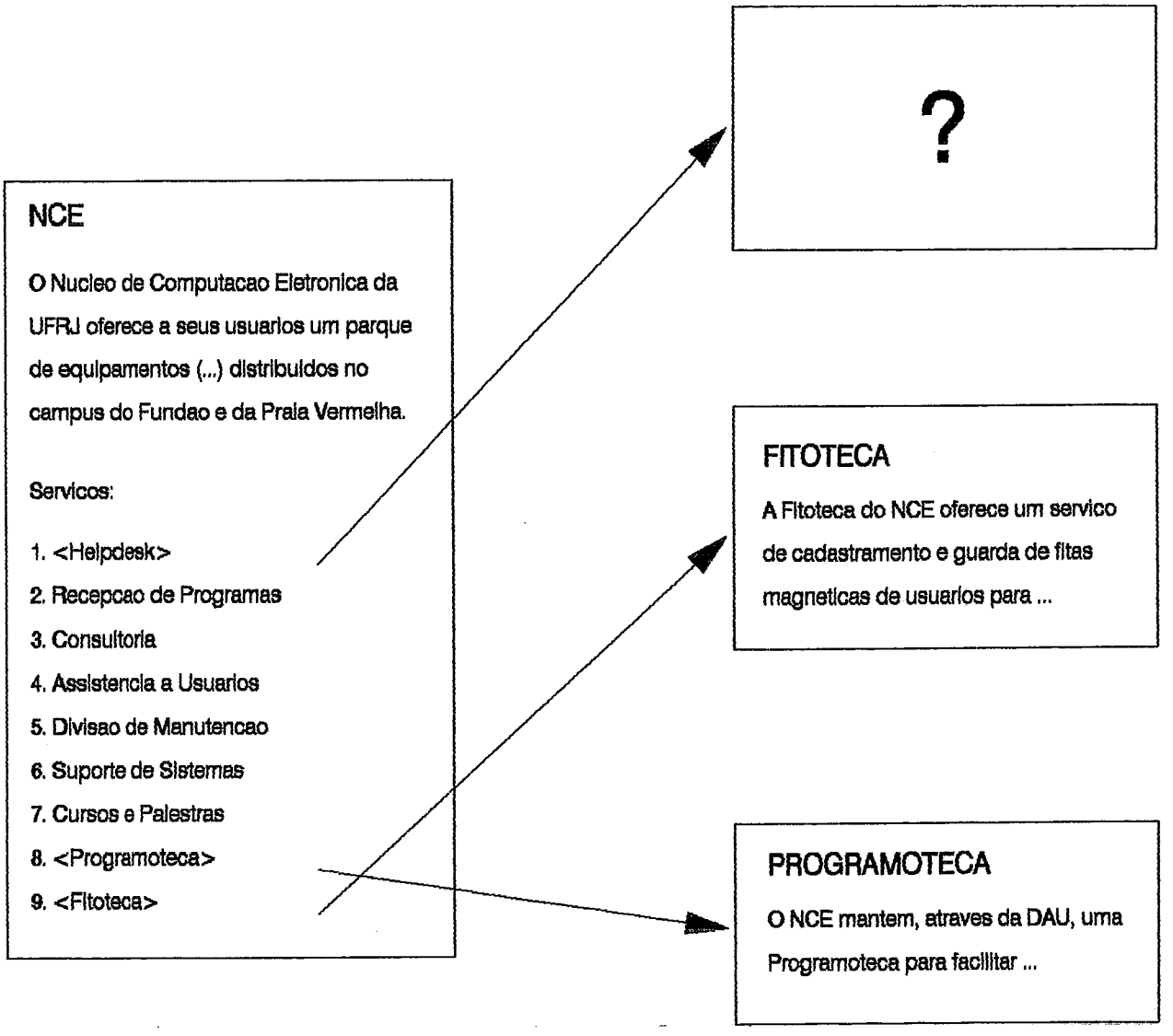


Fig. A1 - Qual o destino de uma ligação que parte do botão "Helpdesk"?

Ligações foram aqui consideradas como unidirecionais e portanto potencialmente elas partem de uma relação (ou subconjunto) "fraca" (baixo poder de resolução) para nós onde essas mesmas relações são "fortes" (alto poder de resolução).

Assim, ao escolhermos o botão **Helpdesk**, em "NCE", automaticamente o sistema buscaria na tabela de ligações (Fig. A6) - que por sua vez contém todas as relações exportadas pelos nós -, quais os nós que, provavelmente, serviriam como destino das ligações que partiriam daquele botão. Ou seja, quais os nós que possuem o termo "helpdesk" contido no conjunto exportado (Fig A2). Ficaria a critério do usuário/autor confirmar, ou não, o estabelecimento dessa ligação.

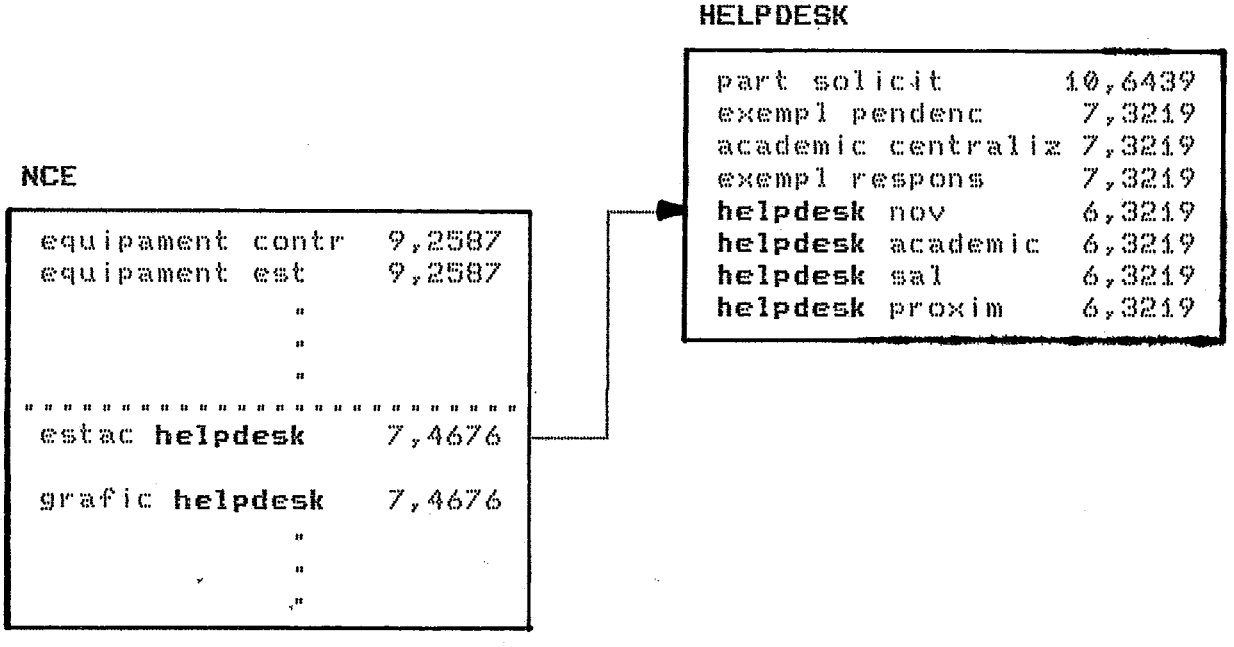


Fig. A2 - Ligações que partem do nó corrente

Uma outra questão surge quando criamos um nó e não sabemos se o conteúdo desse nó já foi incluído em um outro nó já inserido na rede. Por exemplo, ao inserirmos um nó "FITAS" não se tem certeza da existência de um outro nó que se refere basicamente ao mesmo assunto (Fig. A3).

Os nós contidos na rede devem diferir de alguma forma para que um deles possa complementar as idéias contidas no outro e conseqüentemente evitar leituras desnecessárias e tudo aquilo que delas decorre.

Inicialmente, ao inserirmos o nó "FITOTECA", pesos foram atribuídos às relações extraídas, e o conjunto da Fig. A4 foi gerado. Posteriormente ao inserir o nó "FITAS" foi observado uma interseção dos conjuntos exportados pelos nós, indicando que, provavelmente, tais nós continham a mesma idéia. Esses nós são então, sérios candidatos a análise para um possível acoplamento, ou até mesmo para que um deles seja eliminado. No caso, o nó "FITAS" foi eliminado.

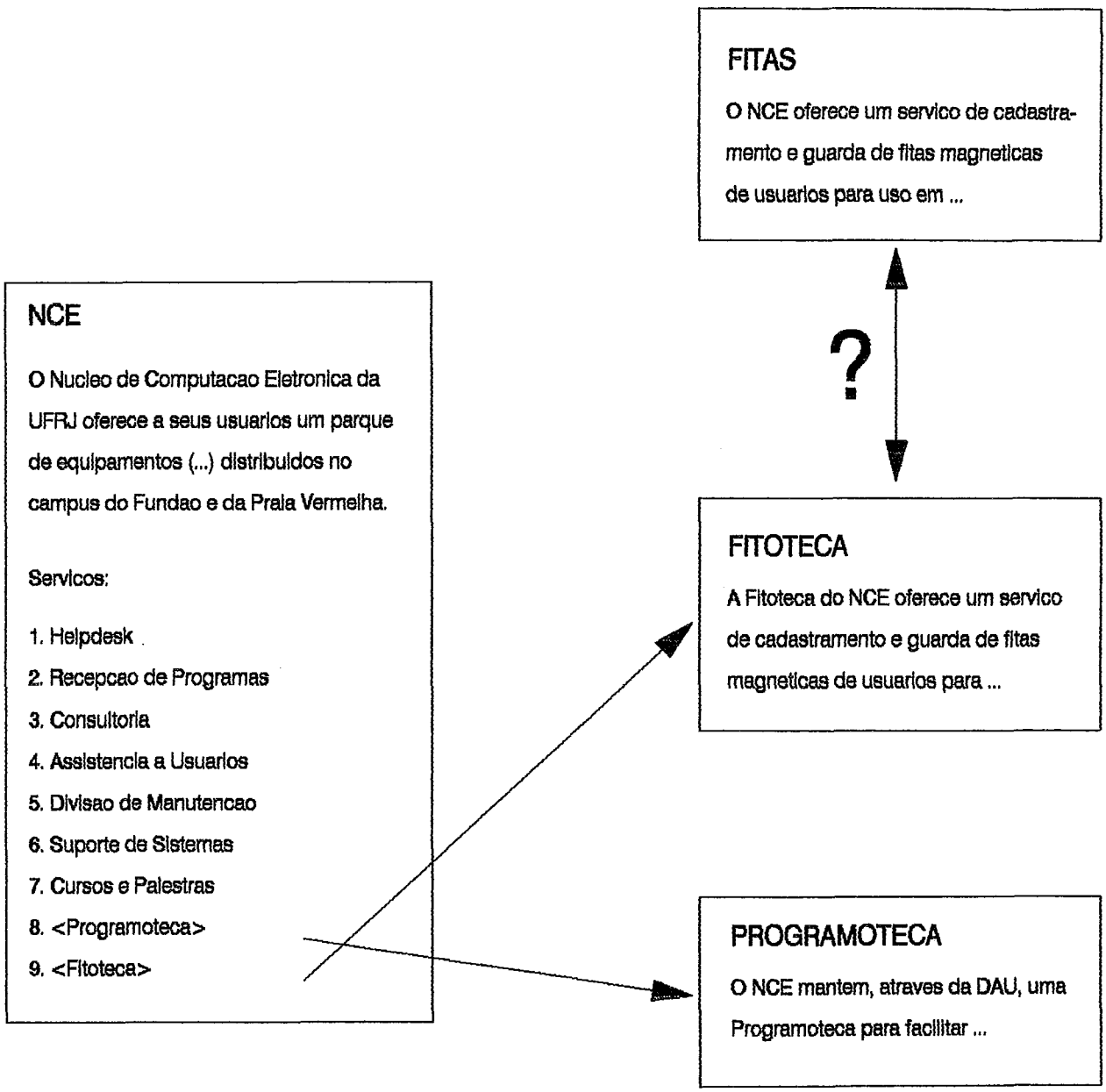


Fig. A3 - Quais nós similares já existem na rede?

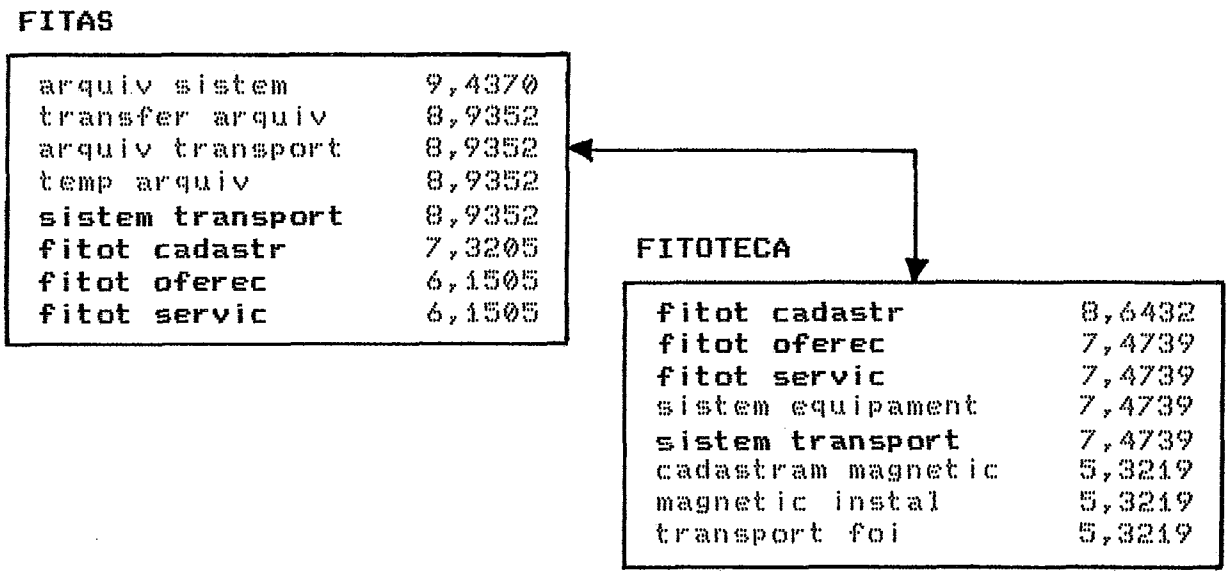


Fig. A4 - Prováveis relacionamentos entre conjuntos de relações léxicas extraídas dos nós.

Finalmente, quais nós, já existentes, devem apontar para o nó "NCE" que está sendo inserido? (caso c, Cap. VII). Certamente esse é um processo um pouco mais complicado pois exige uma reanálise dos nós já inseridos na rede para que possamos nos certificar da existência de um botão que servirá como origem de uma ligação que apontará para esse novo nó (Fig. A5).

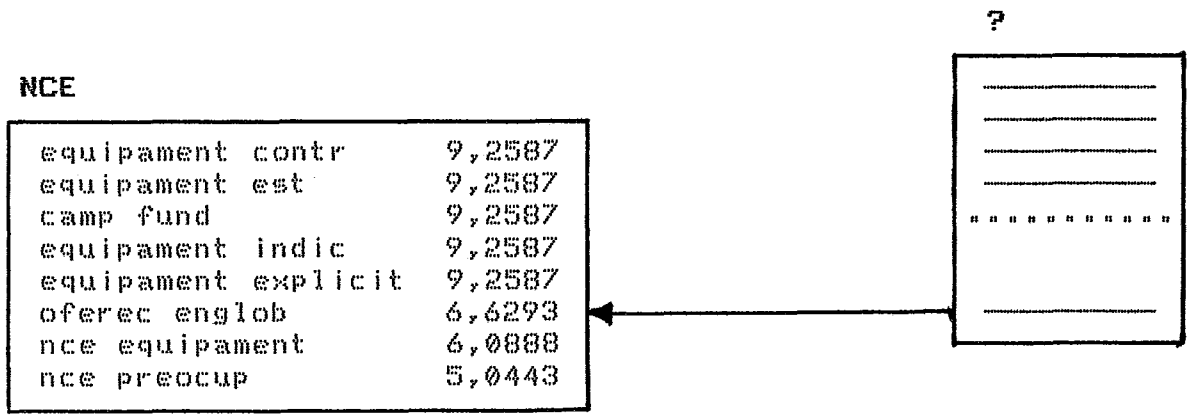


Fig. A5 - Ligações que chegam ao nó corrente.

```
nce        equipament contr
nce        equipament est
nce        camp fund
nce        equipament indic
nce        equipament explicit
nce        oferec englob
nce        nce equipament
nce        nce preocup
fitas      arquiv sistem
fitas      transfer arquiv
fitas      arquiv transport
fitas      temp arquiv
fitas      sistem transport      fitoteca
fitas      fitot cadastr         fitoteca
fitas      fitot oferec          fitoteca
fitas      fitot servic          fitoteca
helpdesk   part solicit
helpdesk   exempl pendenc
helpdesk   academic centraliz
helpdesk   exempl respons
helpdesk   helpdesk nov
helpdesk   helpdesk academic
helpdesk   helpdesk sal
helpdesk   helpdesk proxim
fitoteca   fitot cadastr         fitas
fitoteca   fitot oferec          fitas
fitoteca   fitot servic          fitas
fitoteca   sistem equipament
fitoteca   sistem transport      fitas
fitoteca   cadastram magnetic
fitoteca   magnetic inst
fitoteca   transport foi
```

Fig. A6 - Exemplo de uma Tabela de Ligações

Apêndice B

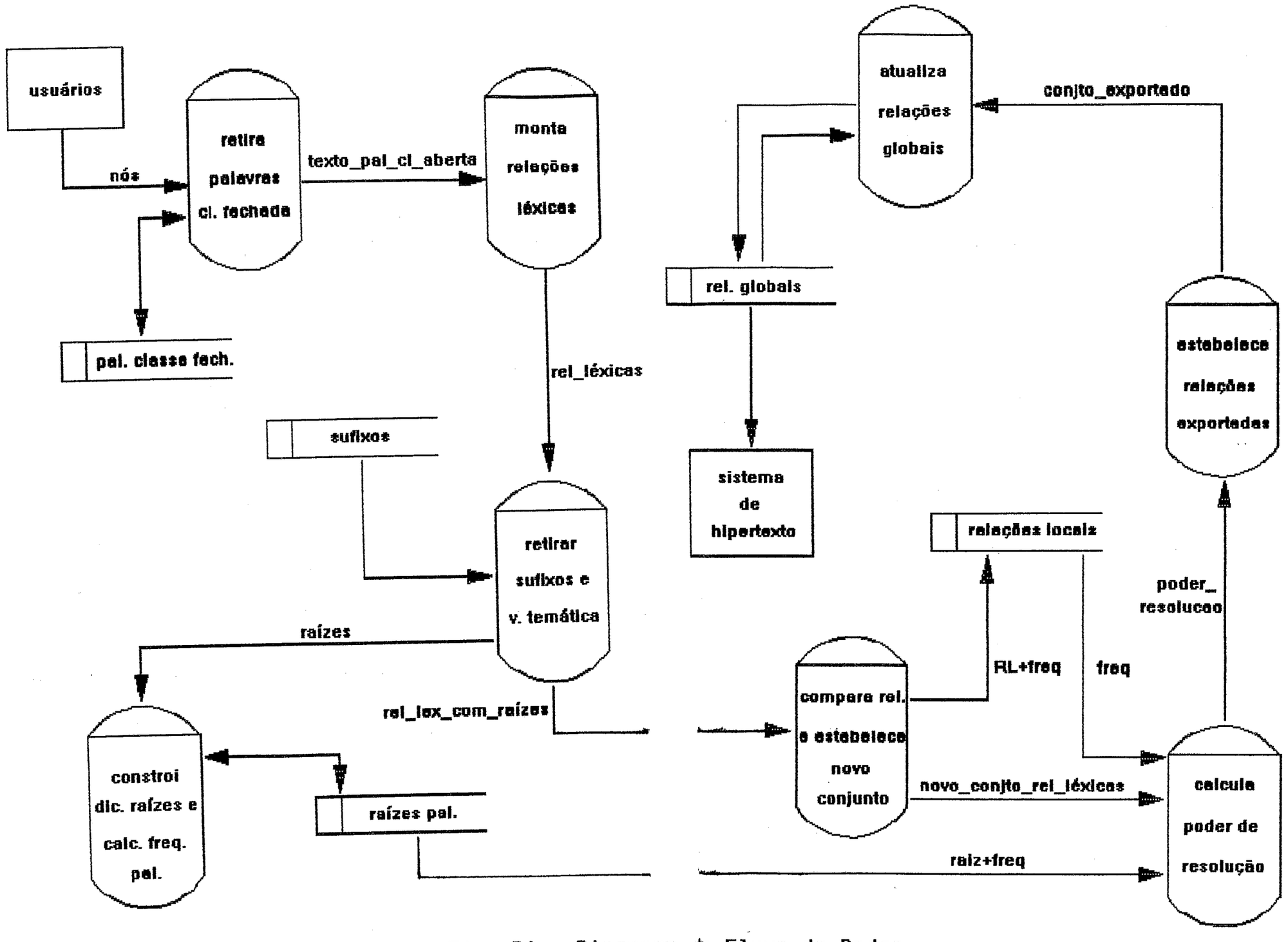


Fig. B1 - Diagrama de Fluxo de Dados